**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**VIỆN KỸ THUẬT CÔNG NGHỆ**

**\*\*\*\*\*\*\*\*\*\*\***



**BÁO CÁO CHUYÊN ĐỀ**

**THỰC TẬP DOANH NGHIỆP**

**ĐỀ TÀI: PHÂN TÍCH VÀ DỰ ĐOÁN KHÁCH HÀNG NGƯNG SỬ DỤNG DỊCH VỤ VIỄN THÔNG**

**Giảng viên hướng dẫn: TS**.**Nguyễn Thị Hồng**

**Sinh viên thực hiện: Cao Thành Phát**

**MSSV: 2124802010172**

**Lớp: D21CNTT04**

**Chuyên ngành: Công Nghệ Thông Tin**

**Khóa: 2021-2026**

**Bình Dương, tháng 07/2024**

**2021 - 2026**

UBND TỈNH BÌNH DƯƠNG

**TRƯỜNG ĐẠI HỌC THỦ DẦU MỘT**

**Cao Thành Phát**

BÁO CÁO CHUYÊN ĐỀ

THỰC TẬP DOANH NGHIỆP

Nơi thực tập

**TRUNG TÂM CNTT THUỘC VIỄN THÔNG BÌNH DƯƠNG**

**TÊN ĐỀ TÀI:**

**D21CNTT04**

**PHÂN TÍCH VÀ DỰ ĐOÁN KHÁCH HÀNG NGƯNG SỬ DỤNG DỊCH VỤ VIỄN THÔNG**

NGƯỜI HƯỚNG DẪN TẠI ĐƠN VỊ: Nguyễn Minh Thiện

GIẢNG VIÊN HƯỚNG DẪN: TS. Lê Thị Hồng

SINH VIÊN THỰC HIỆN: Cao Thành Phát MÃ SỐ SV: 2124802010172

CHUYÊN NGÀNH: Công Nghệ Thông Tin LỚP: D21CNTT04

NIÊN KHÓA: 2021 - 2026

**GIẤY TIẾP NHẬN SINH VIÊN THỰC TẬP**

**PHIẾU NHẬN XÉT SINH VIÊN**

**CỦA ĐƠN VỊ THỰC TẬP**

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

# LỜI CẢM ƠN

Trong thời gian 2 tháng thực tập tại Công ty Viễn thông Bình Dương, em đã có cơ hội trải nghiệm một môi trường làm việc chuyên nghiệp, thân thiện và đầy niềm vui. Những kiến thức và kinh nghiệm thực tế mà em đã học được trong suốt thời gian này vô cùng quý báu, giúp em vững vàng hơn trong việc chuẩn bị cho tương lai nghề nghiệp của mình. Em xin chân thành cảm ơn ban giám đốc đã tạo điều kiện thuận lợi và các anh chị trong công ty đã tận tình hướng dẫn và hỗ trợ em.

Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến cô Nguyễn Thị Hồng cùng toàn thể thầy cô trong viện Kỹ Thuật – Công Nghệ đã truyền đạt những kiến thức quan trọng, giúp em có thể hoàn thành tốt đợt thực tập này.

Vì kiến thức bản thân còn hạn chế, trong quá trình thực tập, hoàn thiện báo cáo chuyên đề này em không tránh khỏi những sai sót, kính mong nhận được những ý kiến đóng góp từ cô cũng như quý công ty.

Em xin chân thành cảm ơn.

# MỤC LỤC

[LỜI CẢM ƠN 5](#_Toc172787945)

[MỤC LỤC 6](#_Toc172787946)

[DANH MỤC HÌNH 8](#_Toc172787947)

[DANH MỤC BẢNG 9](#_Toc172787948)

[MỞ ĐẦU 1](#_Toc172787949)

[1. Lý do thực tập 1](#_Toc172787950)

[2. Mục tiêu thực tập 1](#_Toc172787951)

[3. Địa điểm thực tập 1](#_Toc172787952)

[4. Bố cục bài báo cáo 1](#_Toc172787953)

[CHƯƠNG 1. TỔNG QUAN VỀ VIỄN THÔNG BÌNH DƯƠNG 3](#_Toc172787954)

[1.1 Thành lập và phát triển 3](#_Toc172787955)

[1.2 Sơ đồ tổ chức và bố trí nhân sự 3](#_Toc172787956)

[1.3 Ngành Nghề kinh doanh chính 4](#_Toc172787957)

[1.4 Sứ mệnh và tầm nhìn 4](#_Toc172787958)

[CHƯƠNG 2. NỘI DUNG THỰC TẬP 6](#_Toc172787959)

[2.1 GIỚI THIỆU ĐỀ TÀI 6](#_Toc172787960)

[**2.1.1 Tên đề tài** 6](#_Toc172787961)

[**2.1.2 Một số công việc đã làm được** 6](#_Toc172787962)

[2.2 CÔNG NGHỆ SỬ DỤNG 6](#_Toc172787963)

[**2.2.1 Oracle** 6](#_Toc172787964)

[**2.2.2 Power BI** 7](#_Toc172787965)

[**2.2.3 Giới thiệu về Visual Studio Code** 8](#_Toc172787966)

[2.3 MÔ TẢ BỘ DỮ LIỆU 10](#_Toc172787967)

[**2.4** **PHÂN TÍCH VÀ TRỰC QUAN HOÁ DỮ LIỆU VỚI POWER BI** 12](#_Toc172787968)

[**2.4.1 Tổng quan về bảng báo cáo** 12](#_Toc172787969)

[**2.4.2 Phân tích khả năng khách hàng rời bỏ theo đối tượng** 13](#_Toc172787970)

[**2.4.3 Phân tích khả năng khách hàng rời bỏ dựa trên mức độ hài lòng của khách hàng** 14](#_Toc172787971)

[**2.4.4 Phân tích khả năng khách hàng rời bỏ dựa trên số lần báo hỏng** 15](#_Toc172787972)

[**2.4.5 Phân tích khả năng khách hàng rời bỏ theo thời gian sử dụng dịch vụ** 15](#_Toc172787973)

[**2.4.6 Phân tích khả năng khách hàng rời bỏ theo loại dịch vụ** 16](#_Toc172787974)

[**2.4.7 Phân tích khả năng khách hàng rời bỏ theo tổng nợ** 17](#_Toc172787975)

[**2.4.8 Phân tích khả năng khách hàng rời bỏ theo trả trước** 18](#_Toc172787976)

[2.5 DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ 20](#_Toc172787977)

[**2.5.1 Xử lý dữ liệu** 20](#_Toc172787978)

[**2.5.2 Chuẩn hoá bộ dữ liệu** 25](#_Toc172787979)

[**2.5.3 Xây dựng mô hình Machine Learning** 28](#_Toc172787980)

[**2.5.4 Giới thiệu về thuật toán Decision Tree** 33](#_Toc172787981)

[**2.5.5 Kết quả mô hình Decision Tree** 33](#_Toc172787982)

[**2.5.6 Kết luận** 39](#_Toc172787983)

[**2.5.7 Đề xuất cải tiến** 39](#_Toc172787984)

[CHƯƠNG 3. ĐÁNH GIÁ KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN 40](#_Toc172787985)

[3.1. ĐÁNH GIÁ KẾT QUẢ 40](#_Toc172787986)

[**3.1.1 Kết quả đã làm được** 40](#_Toc172787987)

[**3.1.2 Hạn chế** 40](#_Toc172787988)

[3.2 HƯỚNG PHÁT TRIỂN 41](#_Toc172787989)

[3.3 KẾT LUẬN 41](#_Toc172787990)

[TÀI LIỆU THAM KHẢO 42](#_Toc172787991)

# DANH MỤC HÌNH

[**Hình 1.1 - Logo của công ty 3**](#_Toc172712341)

[**Hình 2.2 - Oracle Database 8**](#_Toc172712342)

[**Hình 2.3 - Power BI Desktop 10**](#_Toc172712343)

[**Hình 2.1 - Visual Studio Code 11**](#_Toc172712344)

[**Hình 2.5 - Bảng báo cáo bộ dữ liệu tổng quan 14**](#_Toc172712345)

[**Hình 2.6 - Bảng báo cáo phân tích khả năng rời bỏ theo đối tượng 15**](#_Toc172712346)

[**Hình 2.7 - Phân tích khả năng khách hàng rời bỏ theo mức độ hài lòng 16**](#_Toc172712347)

[**Hình 2.8 - Phân tích khả năng khách hàng rời bỏ dựa trên số lần báo hỏng 17**](#_Toc172712348)

[**Hình 2.9 - Phân tích khả năng khách hàng rời bỏ theo thời gian sử dụng dịch vụ 18**](#_Toc172712349)

[**Hình 2.10 - Phân tích khả năng khách hàng rời bỏ theo loại dịch vụ 19**](#_Toc172712350)

[**Hình 2.11 - Phân tích khả năng khách hàng rời bỏ dịch vụ theo tổng nợ 20**](#_Toc172712351)

[**Hình 2.12 - Phân tích khả năng khách hàng rời bỏ theo trả trước 21**](#_Toc172712352)

[**Hình 2.13 - Thông tin về bộ dữ liệu 22**](#_Toc172712353)

[**Hình 2.14 - Xử lý giá trị null cột SOTHANG\_DATCOC 23**](#_Toc172712354)

[**Hình 2.15 - Xử lý dữ liệu null cột TRATRUOC 23**](#_Toc172712355)

[**Hình 2.16 - Xử lý dữ liệu null với cột TONGNO 23**](#_Toc172712356)

[**Hình 2.17 - Mô tả bộ dữ liệu của Tenure\_month 23**](#_Toc172712357)

[**Hình 2.18 - Xử lý giá trị null cột Tenure\_month 23**](#_Toc172712358)

[**Hình 2.19 - Các giá trị của cột TOCDO 24**](#_Toc172712359)

[**Hình 2.20 - Kết quả sau khi xử lý cột TOCDO 26**](#_Toc172712360)

[**Hình 2.21 - Kết quả sau khi làm xong các bước xử lý dữ liệu 27**](#_Toc172712361)

[**Hình 2.22 - Biểu đồ mối quan hệ tương quan giữa các biến 28**](#_Toc172712362)

[**Hình 2.23 - Biến đầu vào(X) 29**](#_Toc172712363)

[**Hình 2.24 - Biến mục tiêu (Y) 29**](#_Toc172712364)

[**Hình 2.25 - Giá trị biến mục tiêu 29**](#_Toc172712365)

[**Hình 2.26 - Sử dụng Standard Scaler chuẩn hoá bộ dữ liệu 30**](#_Toc172712366)

[**Hình 2.27 - Kết quả thực hiện mô hình huấn luyện 32**](#_Toc172712367)

[**Hình 2.28 - Kết quả sau khi dùng cross-validation score 33**](#_Toc172712368)

[**Hình 2.29 - So sánh độ chính xác của các mô hình 34**](#_Toc172712369)

[**Hình 2.31 - Mô hình Decision Tree 35**](#_Toc172712370)

[**Hình 2.32 - Ma trận hỗn hợp 36**](#_Toc172712371)

[**Hình 2.33 - Công thức tính điểm Precision và recall trong bài toán phân loại 37**](#_Toc172712372)

[**Hình 2.34 - Ma trận nhầm lẫn chuẩn hóa 38**](#_Toc172712373)

[**Hình 2.35 - Bảng báo cáo phân loại 39**](#_Toc172712374)

# DANH MỤC BẢNG

[**Bảng 2.1. Các trường thông tin của bộ dữ liệu** 12](#_Toc172712631)

[**Bảng 2.2 . Các trường thông tin cần thiết** 13](#_Toc172712632)

# MỞ ĐẦU

## 1. Lý do thực tập

Ngoài việc học những kiến thức trên trường, thực tập còn là cơ hội quý báu để em áp dụng những lý thuyết đã học vào thực tế, giúp em hiểu rõ hơn về ngành nghề mình đang theo đuổi. Thực tập không chỉ giúp em làm quen với môi trường làm việc chuyên nghiệp mà còn rèn luyện các kỹ năng mềm cần thiết như kỹ năng giao tiếp, làm việc nhóm, quản lý thời gian và giải quyết vấn đề. Đây là những kỹ năng mà em khó có thể nắm bắt một cách trọn vẹn chỉ qua sách vở và lý thuyết trên giảng đường.

Thông qua việc thực tập, em có cơ hội tiếp cận với những công nghệ mới, những quy trình làm việc hiện đại, và những phương pháp giải quyết vấn đề sáng tạo. Điều này không chỉ làm phong phú thêm kiến thức chuyên môn mà còn giúp em cập nhật và nắm bắt kịp thời những xu hướng phát triển trong ngành. Hơn nữa, trải qua quá trình thực tập, em sẽ tích lũy được những kinh nghiệm thực tiễn quý giá, chuẩn bị cho công việc sau khi tốt nghiệp một cách tự tin và vững vàng hơn.

Thực tập cũng là cơ hội để em xây dựng và mở rộng mạng lưới quan hệ với các đồng nghiệp và chuyên gia trong ngành. Những mối quan hệ này có thể mang lại nhiều lợi ích lâu dài, không chỉ trong việc tìm kiếm cơ hội việc làm mà còn trong việc học hỏi và phát triển nghề nghiệp. Khi tiếp xúc và làm việc cùng những người đã có nhiều kinh nghiệm, em sẽ học hỏi được nhiều điều bổ ích và có cái nhìn thực tế hơn về ngành nghề mà mình theo đuổi.

Ngoài ra, trong quá trình thực tập, em cũng có thể nhận được những phản hồi, đánh giá từ các đồng nghiệp và cấp trên, giúp em nhận ra những điểm mạnh và điểm yếu của bản thân. Từ đó, em có thể hoàn thiện và phát triển mình một cách toàn diện hơn, không chỉ về mặt chuyên môn mà còn về mặt kỹ năng và phẩm chất cá nhân.

## 2. Mục tiêu thực tập

Mục tiêu thực tập của em là áp dụng những kiến thức lý thuyết đã học vào thực tiễn, giúp em hiểu rõ hơn về ngành nghề mình đang theo đuổi. Em mong muốn phát triển các kỹ năng chuyên môn thông qua việc tham gia vào các dự án và công việc cụ thể, từ đó nâng cao khả năng thực hiện công việc một cách hiệu quả. Bên cạnh đó, em đặt mục tiêu rèn luyện các kỹ năng mềm như giao tiếp, làm việc nhóm, quản lý thời gian và giải quyết vấn đề. Việc xây dựng mối quan hệ với các đồng nghiệp, chuyên gia trong ngành cũng là một trong những ưu tiên của em, nhằm học hỏi kinh nghiệm và mở rộng mạng lưới quan hệ. Em cũng mong muốn tiếp cận và học hỏi về các công nghệ mới, các quy trình làm việc hiện đại để nắm bắt kịp thời những thay đổi và xu hướng mới trong lĩnh vực chuyên môn. Qua quá trình thực tập, em hy vọng sẽ tích lũy được kinh nghiệm thực tiễn quý giá, từ đó có cái nhìn toàn diện hơn về những thách thức và cơ hội trong ngành. Em cũng đặt mục tiêu phát triển bản thân một cách toàn diện, nhận diện và cải thiện điểm mạnh, điểm yếu của mình thông qua phản hồi từ đồng nghiệp và cấp trên. Cuối cùng, em mong muốn chuẩn bị tốt nhất cho công việc sau khi tốt nghiệp, tự tin bước vào thị trường lao động với hành trang kinh nghiệm và kỹ năng cần thiết, xác định rõ định hướng nghề nghiệp và xây dựng kế hoạch phát triển bản thân trong tương lai.

## 3. Địa điểm thực tập

Công ty Viễn thông Bình Dương

Địa chỉ: 326 Đại lộ Bình Dương, Phú Hoà, Thủ Dầu Một, Bình Dương.

## 4. Bố cục bài báo cáo

Nội dung báo cáo gồm 3 phần:

Chương 1: Tổng quan về Công Ty Viễn thông Bình Dương

Chương 2: Nội dung thực tập

Chương 3: Đánh giá kết quả và hướng phát triển

# CHƯƠNG 1. TỔNG QUAN VỀ VIỄN THÔNG BÌNH DƯƠNG

## Thành lập và phát triển



Hình 1.1 - Logo của công ty

Viễn thông Bình Dương được thành lập theo quyết định số 605/QĐ-TCCB-HĐQT ngày 06/12/2007 của Hội đồng Quản trị Tập đoàn Bưu chính Viễn thông Việt Nam trên cơ sở tổ chức lại các đơn vị kinh doanh dịch vụ viễn thông, công nghệ thông tin và các đơn vị trực thuộc khác của Bưu điện tỉnh Bình Dương (cũ), và chính thức đi vào hoạt động kể từ ngày 01/01/2008.

Viễn thông Bình Dương là một doanh nghiệp nhà nước, có bề dày truyền thống phục vụ và kinh doanh các dịch vụ viễn thông – công nghệ thông tin, có hệ thống hạ tầng rộng khắp, đảm bảo phục vụ nhu cầu thông tin của tổ chức Đảng và cơ quan chính quyền địa phương và đủ sức đáp ứng nhu cầu sử dụng của quần chúng nhân dân và các doanh nghiệp trong tỉnh.

Trung tâm Kinh doanh trực thuộc VNPT Bình Dương được thành lập từ ngày 01/07/2014, với mô hình là đơn vị kinh tế trực thuộc, hạch toán phụ thuộc VNPT Bình Dương được thành lập theo quyết định số 819/QĐ-TCCBLĐ ngày 02/06/2014 của Tổng Giám đốc Tập đoàn Bưu chính Viễn thông Việt Nam.

Từ ngày 01/10/2015, Trung tâm Kinh doanh VNPT - Bình Dương (viết tắt là VNPT VinaPhone Bình Dương), đơn vị kinh tế trực thuộc Tổng Công ty Dịch vụ Viễn thông được thành lập theo quyết định số 724/QĐ-VNPT-VNP-NS ngày 28/09/2015 của Tổng Giám đốc Tổng Công ty Dịch vụ Viễn thông.

## 1.2 Sơ đồ tổ chức và bố trí nhân sự

- Ban Giám đốc.

* Bà Võ Thị Thanh Hương

Chức vụ: Trưởng Đại diện Tập đoàn VNPT tại Bình Dương - Giám đốc VNPT Bình Dương

* Ông Nguyễn Quốc Mai

Chức vụ: Phó Giám đốc VNPT Bình Dương

* Ông Nguyễn Nhơn Tuấn

Chức vụ: Phó Giám đốc Viễn Thông Bình Dương, Giám đốc Trung tâm Kinh Doanh - VNPT Bình Dương

-  3 phòng chức năng:

* Phòng Kế toán - Kế hoạch.
* Phòng Nhân sự - Tổng hợp.
* Phòng Kỹ Thuật - Đầu tư.

-   9 trung tâm Viễn thông trực thuộc

* Trung tâm Điều hành Thông tin.
* Trung tâm Công nghệ thông tin.
* Trung tâm Viễn thông Thủ Dầu Một.
* Trung tâm Viễn thông Thuận An.
* Trung tâm Viễn thông Dĩ An.
* Trung tâm Viễn thông Tân Uyên.
* Trung tâm Viễn thông Phú Giáo.
* Trung tâm Viễn thông Bến Cát.
* Trung tâm Viễn thông Dầu Tiếng.

**1.3 Ngành nghề kinh doanh chính**

VNPT Bình Dương có chức năng hoạt động sản xuất kinh doanh và phục vụ chuyên ngành Viễn thông - Công nghệ thông tin Cụ thể như sau:

- Kinh doanh các dịch vụ viễn thông, công nghệ thông tin.

- Kinh doanh, cung ứng đại lý vật tư, thiết bị viễn thông, công nghệ thông tin theo yêu cầu sản xuất kinh doanh của đơn vị và nhu cầu của khách hàng.

- Khảo sát, tư vấn lắp đặt, bảo dưỡng các công trình viễn thông, công nghệ thông tin.

**-** Kinh doanh các ngành nghề khác trong phạm vi được Tập đoàn Bưu chính Viễn thông Việt Nam cho phép và phù hợp với quy định của pháp luật.

## 1.4 Sứ mệnh và tầm nhìn

* **Sứ mệnh**

- Cung cấp cho khách hàng và đối tác các sản phẩm, dịch vụ Viễn thông – Công nghệ thông tin - Truyền thông và Dịch vụ số chất lượng, đột phá, sáng tạo.

- Trở thành trung tâm số (Digital Hub) của khu vực châu Á.

- Tôn vinh và đánh giá giá trị đích thực của người lao động.

- Tiên phong trong các hoạt động về cộng đồng.

* Tầm nhìn

- Trở thành nhà cung cấp dịch vụ số hàng đầu Việt Nam vào năm 2025.

- Trở thành Trung tâm số (Digital Hub) của châu Á vào năm 2030.

- Trở thành lựa chọn số 1 của khách hàng sử dụng sản phẩm, dịch vụ Công nghệ thông tin - Viễn thông (ICT) tại thị trường.

# CHƯƠNG 2. NỘI DUNG THỰC TẬP

## 2.1 GIỚI THIỆU ĐỀ TÀI

### **2.1.1 Tên đề tài**

Đề tài “**Phân tích và dự đoán khách hàng ngưng sử dụng dịch vụ viễn thông**”.

### **2.1.2 Một số công việc đã làm được**

* + - * Phân tích dữ liệu khách hàng
* Làm sạch dữ liệu để thực hiện phân tích.
* Khai phá bộ dữ liệu.
* Thực hiện một số phân tích thống kê để xác định các yếu tố ảnh hưởng đến việc ngưng sử dụng dịch vụ.
  + - * Xây dựng mô hình dự đoán khách hang ngưng sử dụng dịch vụ
* Chọn lựa và áp dụng các thuật toán học máy phù hợp
* Chia dữ liệu thành tập huấn luyện và tập kiểm tra để xây dựng và đánh giá mô hình
* Tối ưu hoá mô hình dự đoán để đạt được độ chính xác cao nhất
  + - * Đánh giá và tối ưu hoá mô hình
* Sử dụng các kỹ thuật đánh giá mô hình như cross-validation để đảm bảo mô hình không bị overfitting.
* Điều chỉnh các tham số của mô hình để đạt được hiệu suất tốt nhất.
* So sánh hiệu quả của các mô hình khác nhau và chọn ra mô hình tốt nhất.

## 2.2 CÔNG NGHỆ SỬ DỤNG

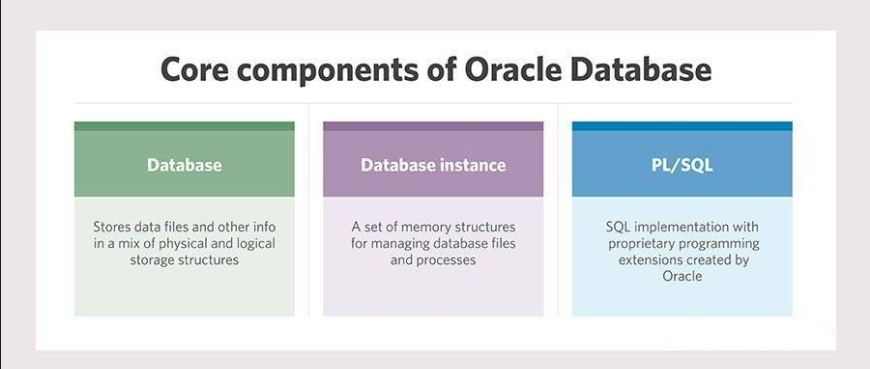
### **2.2.1 Oracle**

Phiên bản hiện đang sử dụng : Oracle Database 12

Oracle là một trong những nhà cung cấp dữ liệu lớn nhất trên thị trường công nghệ hiện nay. Cái tên Oracle được viết tắt từ sản phẩm chủ lực của hãng và hệ thống quản lý cơ sở dữ liệu quan hệ (RDBMS) có tên chính thức là Oracle Database. Phần mềm cơ sở dữ liệu là bộ phận quan trọng và là trung tâm trong mảng IT của nhiều công ty đảm nhiệm nhiều nhiệm vụ khác nhau như xử lý giao dịch, business intelligence (BI) hay các ứng dụng phân tích.

Oracle Database là công nghệ tích hợp và là nền tảng quản lý các dữ liệu được sử dụng nhiều nhất. Nền tảng này đã được sử dụng cho các ứng dụng trên máy tính, các kho chứa hay hệ thống BI.

Oracle Database cũng được xây dựng dựa trên tiêu chuẩn hoá của ngôn ngữ lập trình SQL cho phép quản lý và truy vấn dữ liệu trên máy chủ một cách hiệu quả. Phần mềm hỗ trợ có PL/SQL được Oracle áp dụng để phát triển các tính năng độc quyền cho SQL. Ngoài ra, Oracle Database còn cho phép người dùng lập trình bằng Java, chương trình hoặc ngôn ngữ bất kỳ.



Hình 2.2 - Oracle Database

Oracle Database kết nối phần tử dữ liệu bằng các cấu trúc bảng và cho ra các kết quả lưu trữ khác nhau hỗ trợ xử lý dữ liệu dễ dàng hơn. Các mô hình kiến trúc của Oracle bao gồm một số các ràng buộc ACID đảm bảo tính chính xác cao nhất và độ xử lý tin cậy cho dữ liệu. ACID là các nguyên tắc về nguyên tử, tính thống nhất, độc lập và độ bền của dữ liệu chính.

- Kiến trúc của Oracle Database bao gồm:

+ Cấu trúc lưu trữ vật lý của cơ sở dữ liệu là các tệp chứa dữ liệu, siêu dữ liệu và các tệp quản lý ghi lại thay đổi của dữ liệu. Cơ sở dữ liệu và các phiên bản của nó thực hiện lưu trữ và quản lý các tệp.

+ Cấu trúc lưu trữ logic của Oracle Database bao gồm khối dữ liệu là các phạm vi và nhóm các khối dữ liệu liền kề nhau. Phân đoạn là tập hợp các phạm vi mở rộng. Không gian bảng là các vùng chứa cho phân đoạn.

### ***2.2.2 Power BI***

Phiên bản hiện đang sử dụng: 2.131.1126.0

Power BI Desktop là một ứng dụng miễn phí mà người dùng có thể cài đặt trên máy tính cục bộ cho phép kết nối, chuyển đổi và trực quan hóa dữ liệu của mình. Với Power BI Desktop, người dùng có thể kết nối với nhiều nguồn dữ liệu khác nhau và kết hợp chúng (thường được gọi là mô hình hóa) thành một mô hình dữ liệu cho phép xây dựng hình ảnh và bộ sưu tập hình ảnh mà có thể chia sẻ dưới dạng báo cáo, với những người khác trong tổ chức. Hầu hết người dùng làm việc trong các dự án Business Intelligence sử dụng Power BI Desktop để tạo báo cáo và sau đó sử dụng dịch vụ Power BI để chia sẻ báo cáo của họ với người khác.

Các cách sử dụng phổ biến nhất trong Power BI Desktop:

* Kết nối với dữ liệu :

Bạn có thể kết nối nhiều nguồn dữ liệu khác nhau thông qua Power BI Desktop

* Chuyển đổi và làm sạch dữ liệu đó, để tạo một mô hình dữ liệu :

Trong Power BI Desktop, người dùng có thể dọn dẹp và chuyển đổi dữ liệu bằng Trình soạn thảo truy vấn tích hợp. Với Trình chỉnh sửa truy vấn, bạn có thể thay đổi dữ liệu của mình, chẳng hạn như thay đổi loại dữ liệu, xóa cột hoặc kết hợp dữ liệu từ nhiều nguồn.

* Trực quan hoá dữ liệu bằng các dữ liệu :

Khi bạn có mô hình dữ liệu, bạn có thể kéo các trường vào khung báo cáo để tạo hình ảnh. Một hình ảnh là một đại diện đồ họa của dữ liệu trong mô hình của bạn.

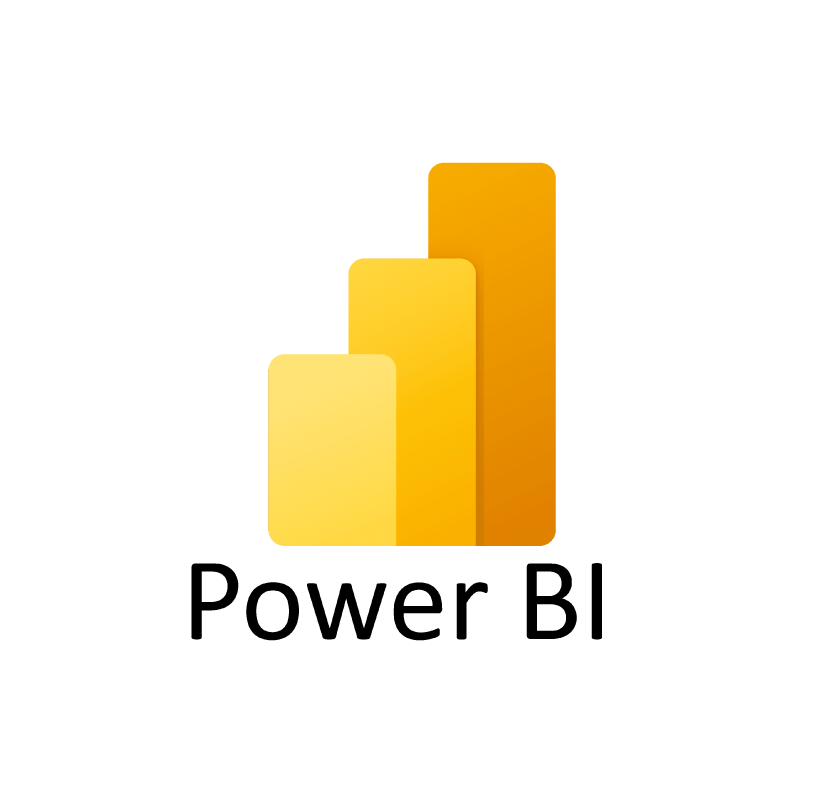
Có nhiều loại hình ảnh khác nhau để lựa chọn trong Power BI Desktop. Để tạo hoặc thay đổi hình ảnh, chỉ cần chọn biểu tượng hình ảnh từ khung Visualization. Nếu bạn có một hình ảnh được chọn trên khung báo cáo, hình ảnh được chọn sẽ thay đổi thành loại bạn đã chọn. Nếu không có hình ảnh nào được chọn, một hình ảnh mới được tạo ra dựa trên lựa chọn của bạn.

* Tạo báo cáo trên 1 hoặc nhiều trang báo cáo :

Một bộ sưu tập hình ảnh, trong một tệp Power BI Desktop, được gọi là báo cáo. Một báo cáo có thể có một hoặc nhiều trang, giống như một tệp Excel có thể có một hoặc nhiều bảng tính

* Chia sẻ báo với người khác bằng dịch vụ PowerBI :

Khi báo cáo đã sẵn sàng để chia sẻ với người khác, bạn có thể Xuất bản báo cáo lên dịch vụ Power BI và cung cấp cho bất kỳ ai trong tổ chức của bạn có giấy phép Power BI.

.

Hình 2.3 - Power BI Desktop

### **2.2.3 Giới thiệu về Visual Studio Code**

Visual Studio Code là một trình soạn thảo, biên tập code hoàn toàn miễn phí dành được Microsoft phát triển cho các lập trình viên và có mặt trên hầu hết các hệ điều hành phổ biến như: Windows, Linux và macOS. Có thể nói rằng, Visual Studio Code là một sự kết hợp độc đáo – đỉnh cao giữa IDE và Code Editor.

Phần mềm này tương thích cực kỳ mạnh mẽ với hệ điều hành Windows. Ngoài ra, bạn còn có thể tải và sử dụng Visual Studio Code trên 2 hệ điều hành vô cùng phổ biến khác là Linux và MacOS.

Không chỉ hỗ trợ cho các ngôn ngữ độc quyền hay được phát triển bởi Microsoft, Visual Studio Code còn có thể hỗ trợ lập trình viên sử dụng nhiều ngôn ngữ lập trình khác nhau thông qua các Extension mở rộng tính năng cho Visual Studio Code.

Một trong những điểm nổi bật khiến Visual Studio Code trở nên vô cùng mạnh mẽ, đó chính là những khả năng như: hỗ trợ hoàn tất cú pháp, tự động canh lề, đối soát dấu ngoặc, …

Visual Studio Code là một phần mềm được Microsoft rất “cưng chiều” cùng sự nhiệt tình ủng hộ cho sự phát triển của Visual Studio Code từ các cộng đồng lập trình viên lớn. Vì thế, Visual Studio Code liên tục được nâng cấp, được trang bị các công nghệ tối ưu nhất và hỗ trợ tốt nhất cho các lập trình viên như:

* Khả năng mở rộng ngôn ngữ
* Intellisense
* Cung cấp kho tiện ích mở rộng khổng lồ
* Hỗ trợ đa nền tảng
* Hỗ trợ tối ưu hóa việc code
* Tối ưu hóa cho việc xây dựng website
* Khả năng tích hợp đầu cuối
* Một màn hình với nhiều chức năng
* Hỗ trợ sử dụng Git



Hình 2.1 - Visual Studio Code

## 2.3 MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu có các trường thông tin như sau:

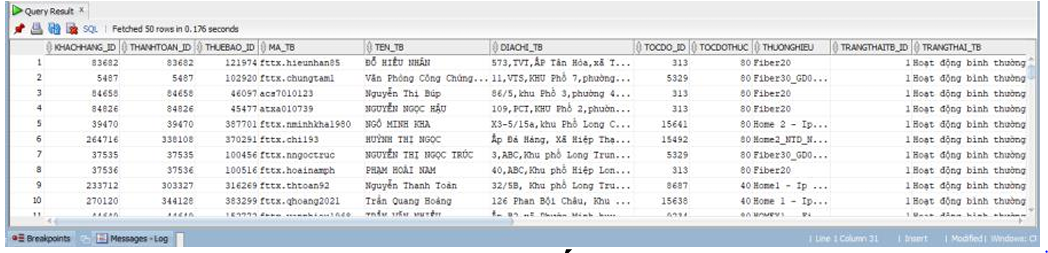
|  |  |
| --- | --- |
| Tên cột | Mô tả |
| THUEBAO\_ID | Mã thuê bao mỗi thuê có 1 mã riêng biệt |
| NGAY\_SD | Thời gian bắt đầu sử dụng của thuê bao |
| NGAYNGUNG\_SD | Thuê bao ngừng sử dụng dịch vụ |
| DICHVU | Dịch vụ thuê bao sử dụng  Fiber,MyTV,MeshWifi |
| THOIHAN | Thời hạn sử dụng dịch vụ |
| TOCDO | Tốc độ của gói cước hoặc dịch vụ |
| TRANGTHAITB | Trạng thái thuê bao |
| TRATRUOC | Thuê bao có trả trước hay không |
| TONGNO | Nợ của thuê bao |
| TRANGBI\_ID | Mã trang bị |
| DOITUONG | Thuê bao thuộc loại đối tượng nào  cn: cá nhân  dn: doanh nghiệp |
| HINHTHUC\_TT | Hình thức thanh toán của thuê bao |
| Tenure\_month | Số tháng thuê bao đã sử dụng dịch vụ |
| SOLANBAOHONG | Số lần báo hỏng của thuê bao trong thời gian sử dụng dịch vụ |
| MUCDOHAILONG | Mức độ hài long của thuê bao |
| TRANGTHAI\_ROIBO | Khách hàng đã rời bỏ  0:Không  1:có |

**Bảng 2.1. Các trường thông tin của bộ dữ liệu**

Các trường thông tin cần thiết cho việc phân tích và xây dựng mô hình dự đoán

|  |  |
| --- | --- |
| Tên cột | Mô tả |
| THUEBAO\_ID | Mã thuê bao mỗi thuê có 1 mã riêng biệt |
| NGAY\_SD | Thời gian bắt đầu sử dụng của thuê bao |
| NGAYNGUNG\_SD | Thuê bao ngừng sử dụng dịch vụ |
| DICHVU | Dịch vụ thuê bao sử dụng  Fiber,MyTV,MeshWifi |
| THOIHAN | Thời hạn sử dụng dịch vụ |
| TOCDO | Tốc độ của gói cước hoặc dịch vụ |
| TRATRUOC | Thuê bao có trả trước hay không |
| TONGNO | Nợ của thuê bao |
| DOITUONG | Thuê bao thuộc loại đối tượng nào  cn: cá nhân  dn: doanh nghiệp |
| Tenure\_month | Số tháng thuê bao đã sử dụng dịch vụ |
| SOLANBAOHONG | Số lần báo hỏng của thuê bao trong thời gian sử dụng dịch vụ |
| MUCDOHAILONG | Mức độ hài long của thuê bao |
| TRANGTHAI\_ROIBO | Khách hàng đã rời bỏ  0:Không  1:có |

**Bảng 2.2 . Các trường thông tin cần thiết**

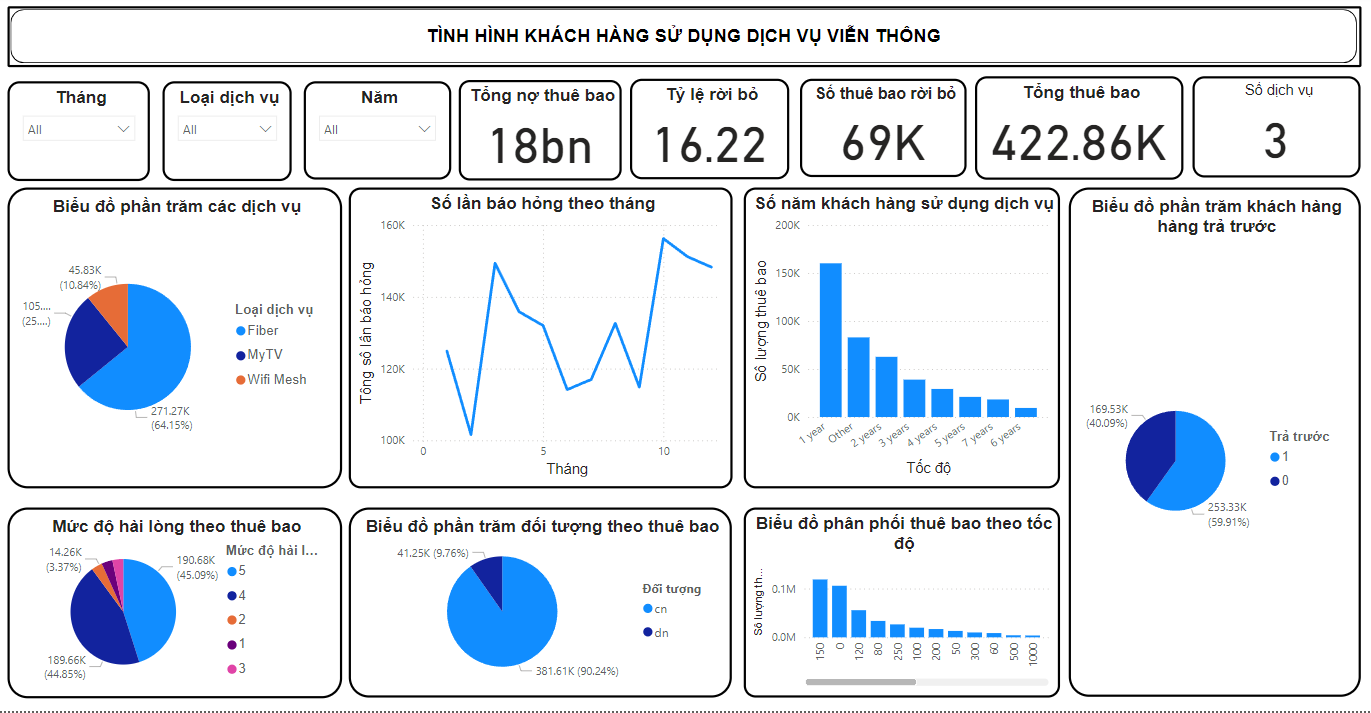
Hình 2.4- Dữ liệu thực tế Oracle

* 1. **PHÂN TÍCH VÀ TRỰC QUAN HOÁ DỮ LIỆU VỚI POWER BI**

### **2.4.1 Tổng quan về bảng báo cáo**

Thiết kế PowerBI về tổng quan các trường thông tin bao gồm:

* Tổng nợ thuê bao
* Tỷ lệ rời bỏ tính theo phần tram
* Số thuê bao rời bỏ
* Tổng thuê bao
* Số dịch vụ
* Biểu đồ phần trăm các dịch vụ (Biểu đồ tròn)
* Số lần báo hỏng theo tháng (Biểu đồ đường)
* Số năm khách hàng sử dụng dịch vụ (Biểu đồ cột)
* Biểu đồ phần trăm khách hàng trả trước (Biểu đồ tròn)
* Biểu đồ phân phối thuê bao theo tốc độ(Biểu đồ cột)
* Biểu đồ phần trăm đối tượng theo thuê bao (Biểu đồ tròn)
* Mức độ hài lòng theo thuê bao (Biểu đồ tròn)
* Các slicer về tháng,năm,loại dịch vụ

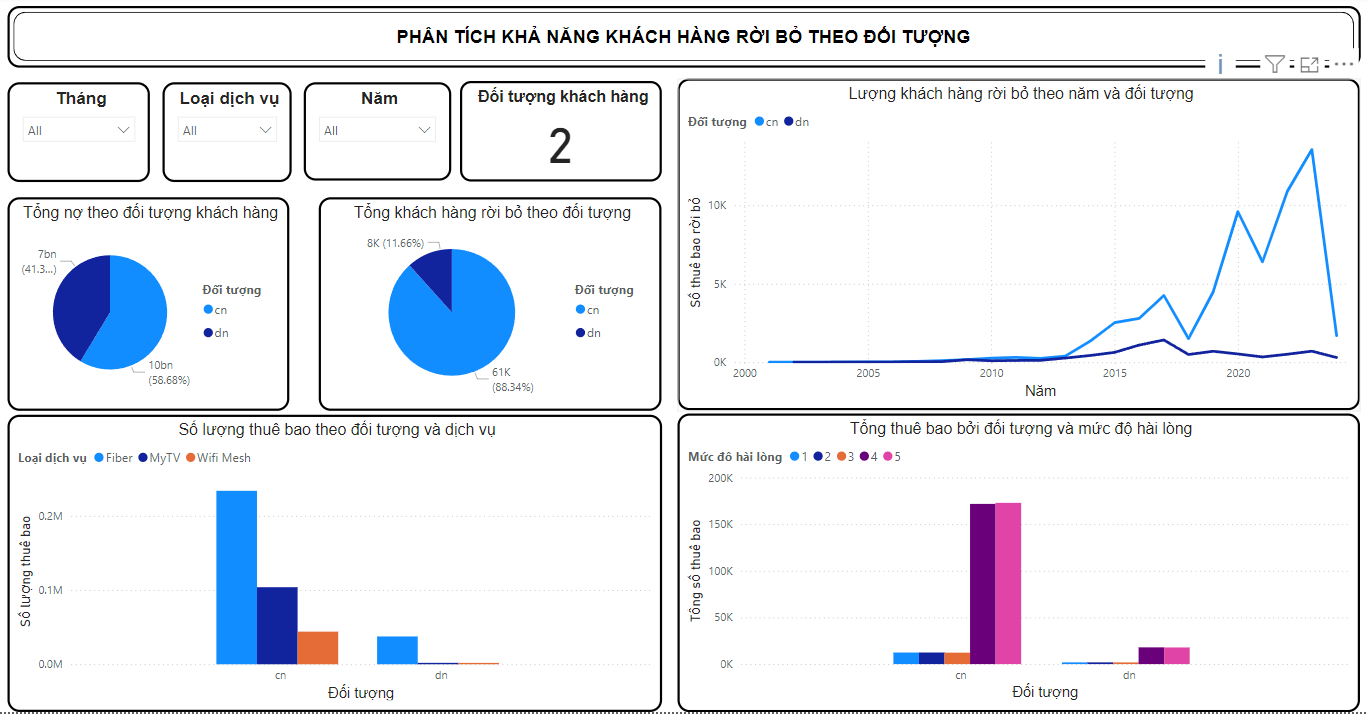


Hình 2.5- Bảng báo cáo bộ dữ liệu tổng quan

Qua bộ dashboard tổng quan trên ta thấy rằng:

* Tổng nợ thuê bao của khách hàng là 18 tỷ.
* Tỷ lệ rời bỏ của khách hàng chiếm 16,22%.
* Số lượng thuê bao rời bỏ khoảng 69 ngàn trên gần 423 ngàn thuê bao.
* Biểu đồ tròn cho chúng ta thấy dịch vụu Fiber có tỷ trọng cao nhất so với các dịch vụ khác.
* Số lần báo theo tháng khá cao đỉnh điểm thường sẽ vào cuối năm.
* Mức độ hài lòng của khách hàng có nhiều thuê bao đánh giá 4 đến 5 điểm so với các điểm còn lại.
* Đối tượng khách hàng chủ yếu là các thuê bao cá nhân chiếm hơn 90%.
* Khách hàng thường sử dụng tốc độ từ 120-150 mpbs.Với những thuê bao có tốc độ 0 mbps đều là những thuê bao sử dụng dịch vụ MyTV
* Gần 60% khách hàng lựa chọn hình thức trả trước.Trong khi hơn 40% khách hàng lựa chọn dịch vụ trả sau.

**2.4.2 Phân tích khả năng khách hàng rời bỏ theo đối tượng**

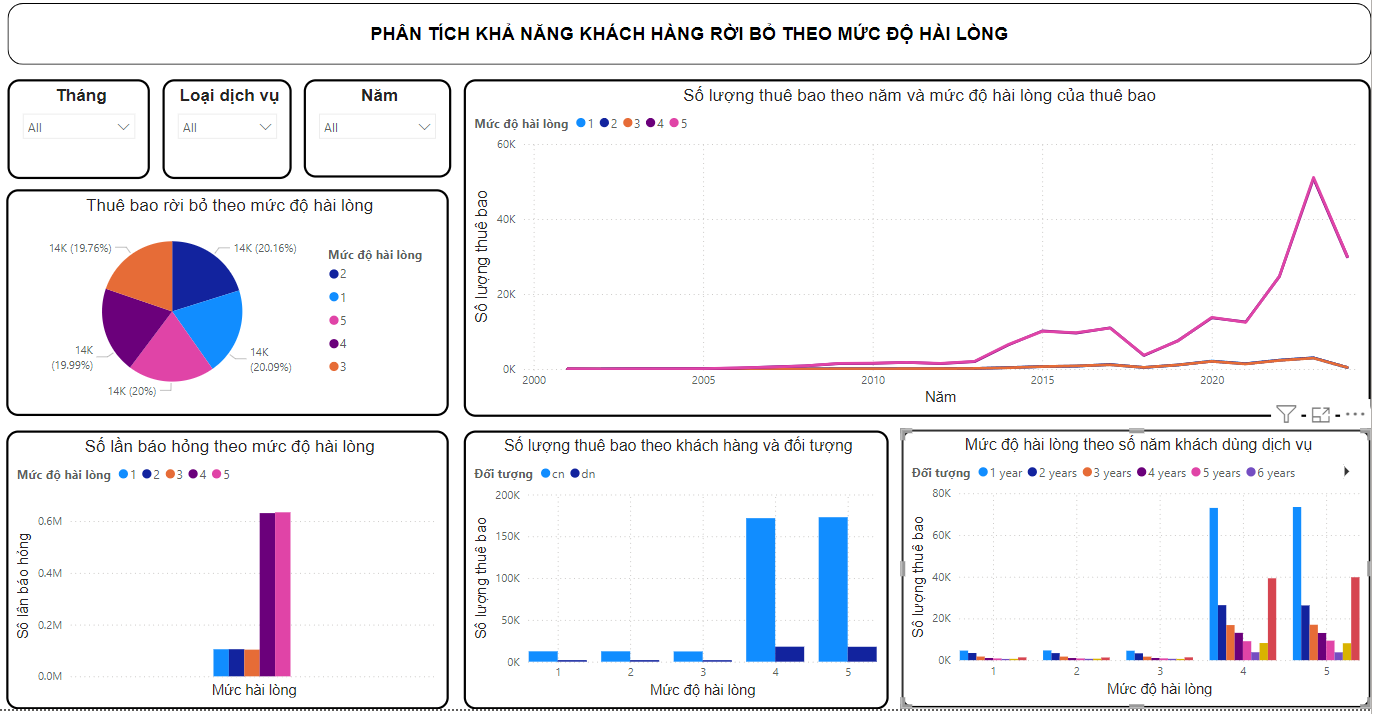


Hình 2.6- Bảng báo cáo phân tích khả năng rời bỏ theo đối tượng

Từ bảng báo cáo trên ta thấy rằng:

* Số khách hàng rời bỏ theo đối tượng cá nhân rất nhiều và tang theo thời gian.
* Tổng nợ của đối tượng cá nhân cao hơn doanh nghiệp .Tuy nhiên, tổng nợ của doanh nghiệp cũng khá cao dù khách hàng doanh nghiệp thấp hơn khách hàng cá nhân.
* Loại dịch vụ mà khách hàng cá nhân và doanh nghiệp sử dụng nhiều nhất là Fiber
* Cả 2 đối tượng đều có mực độ hài lòng cao nhất là 4 và 5 điểm.
* Dựa vào các điểm trên ta có thể rút ra một số kết luận sau:
* Số lượng rời bỏ khách hàng cá nhân tăng theo thời gian do nhiều nguyên nhân do sự cạnh tranh cao của các nhà mạng dịch vụ khác,hoặc những thay đổi trong nhu cầu điều kiện kinh tế của khách hàng.
* Tổng nợ khách hàng cá nhân cao hơn điều này có thể phản ánh xu hướng sử dụng nhiều dịch vụ hơn hay khó khan trong việc thanh toán.Cần có biện pháp kiểm soát nợ và hỗ trợ thanh toán để giảm thiểu rủi ro.
* Dù số lượng doanh nghiệp thấp hơn khách hàng cá nhân ,tổng nợ của doanh nghiệp cũng ở mức cao.Cần có chiến lược quản lý nợ hiệu quả hơn đối với đối tượng này

### **2.4.3 Phân tích khả năng khách hàng rời bỏ dựa trên mức độ hài lòng của khách hàng**



Hình 2.7-Phân tích khả năng khách hàng rời bỏ theo mức độ hài lòng

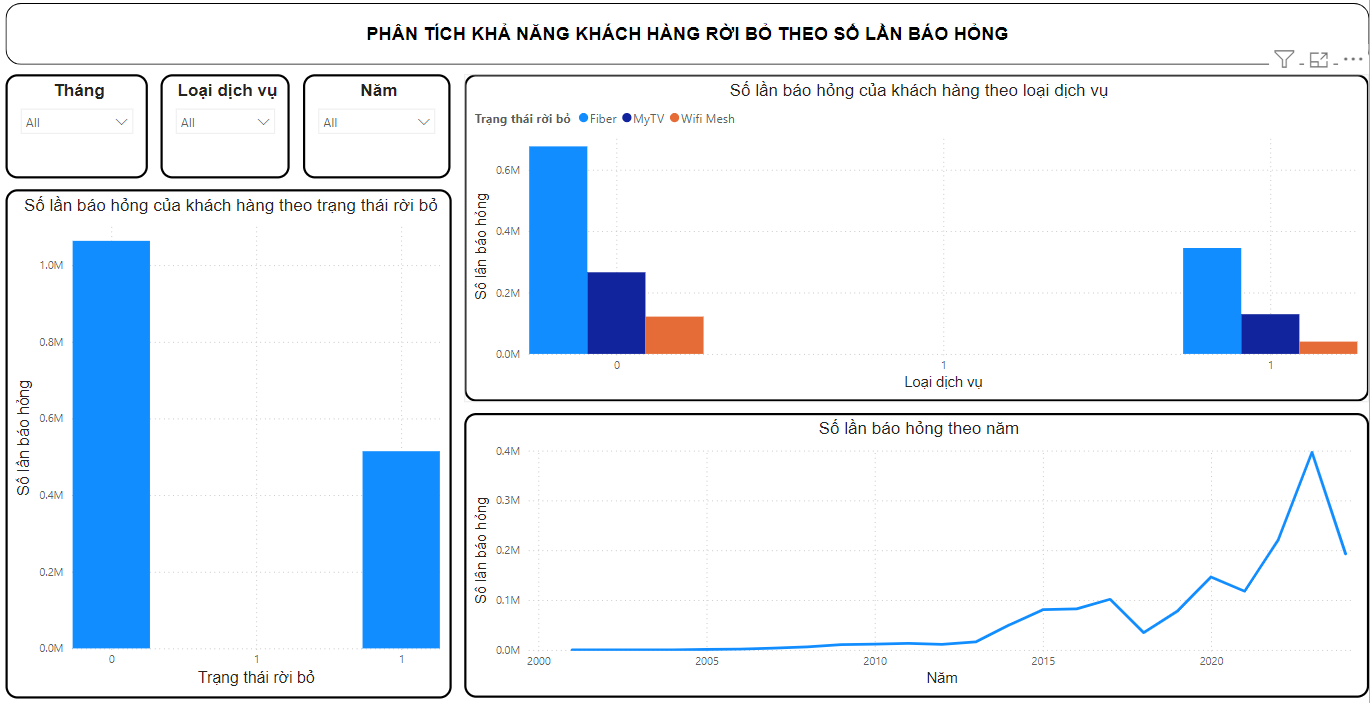
Trong bảng báo cáo “Phân tích khách hàng rời bỏ theo mức độ hài lòng” có mô tả như sau:

* Biểu đồ tròn phân bố thuê bao rời bỏ theo mức độ hài lòng cho thấy tỷ lệ rời bỏ dịch vụ khá đều .Mỗi mức độ hài lòng đều chiếm từ 19-20% tổng số thuê bao rời bỏ ,khồng có mức độ nào quá nổi bật.
* Biểu đồ đường thể hiện số lượng thuê bao theo năm và mức độ hài lòng từ khoảng năm 2020,số lượng thuê bao rời bỏ dịch vụ tang đột biến ,đặc biệt ở mức độ hài lòng 4 và năm.
* Biểu đồ cột cho thấy số lần bảo hỏng tập trung chủ yếu ở mức độ hài lòng 4 và 5.
* Biểu đồ cột cho thấy số lượng thuê bao của khách hàng cá nhân nhiều hơn so với doanh nghiệp ở mức độ hài lòng.
* Biểu đồ cột cho thấy mức độ hài lòng 4 và 5 chiếm ưu thế ở mọi đối tượng ,đặc biệt là những khách hàng đã sử dụng dịch vụ từ 4-5 năm.

Kết luận:

* Mặc dù mức độ hài lòng của khách hàng ở mức 4 và 5 là cao, nhưng vẫn có nhiều thuê bao rời bỏ dịch vụ. Điều này có thể do nhiều nguyên nhân khác ngoài sự hài lòng như giá cả, sự xuất hiện của đối thủ cạnh tranh, hoặc những thay đổi cá nhân của khách hàng.
* Số lượng thuê bao rời bỏ tăng mạnh từ năm 2020, đặc biệt ở những người có mức độ hài lòng cao. Do đó, doanh nghiệp cần điều tra và xác định nguyên nhân cụ thể để có biện pháp khắc phục kịp thời.
* Số lần báo hỏng tập trung ở mức độ hài lòng cao (4 và 5) cho thấy mặc dù khách hàng hài lòng với dịch vụ, họ vẫn gặp nhiều vấn đề kỹ thuật. Cần cải thiện chất lượng kỹ thuật để giảm thiểu số lần báo hỏng và tăng cường trải nghiệm khách hàng.
* Dù mức độ hài lòng 4 và 5 chiếm ưu thế, doanh nghiệp cần tiếp tục duy trì và cải thiện chất lượng dịch vụ để đảm bảo khách hàng tiếp tục hài lòng và giảm tỷ lệ rời bỏ dịch vụ.

### **2.4.4 Phân tích khả năng khách hàng rời bỏ dựa trên số lần báo hỏng**



Hình 2.8-Phân tích khả năng khách hàng rời bỏ dựa trên số lần báo hỏng

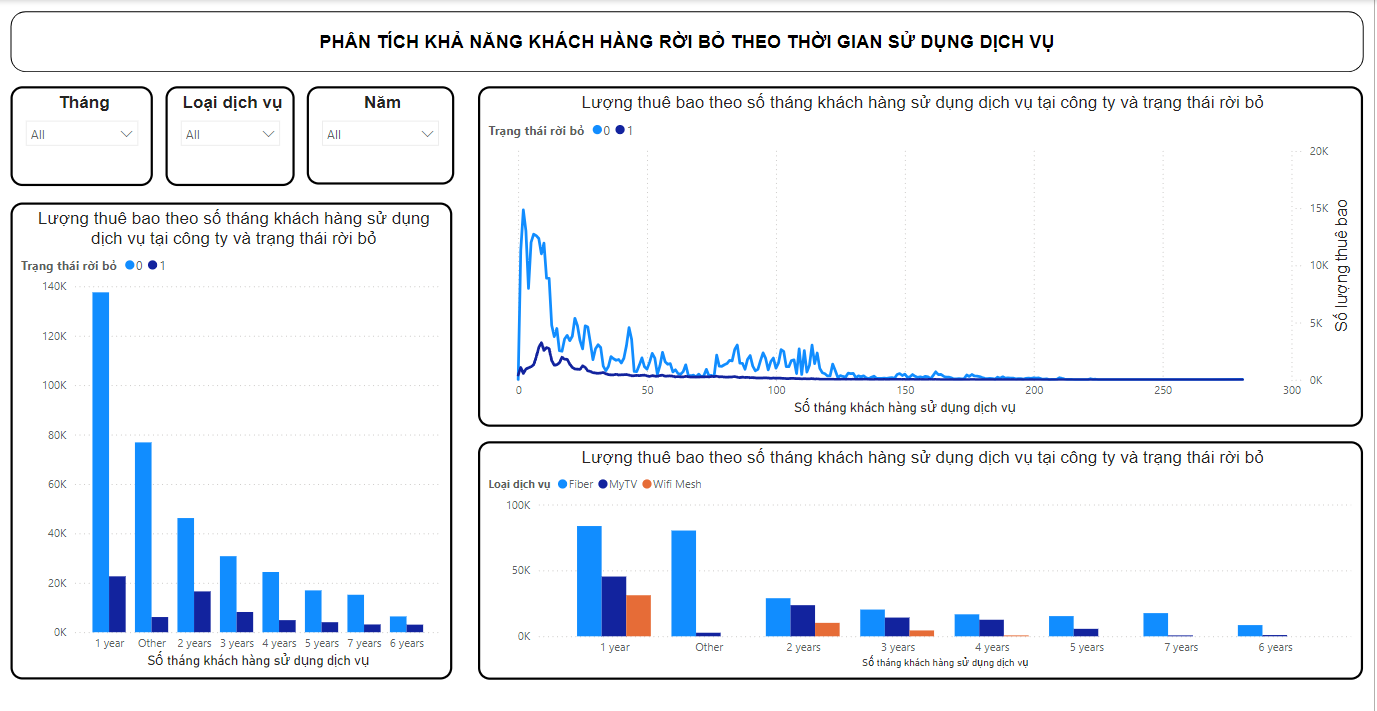
Mô tả:

* Biểu đồ cột cho thấy số lần báo hỏng của khách hàng không rời bỏ dịch vụ cao hơn rất nhiều so với những khách hàng đã rời bỏ dịch vụ.
* Biểu đồ cột cho thấy loại dịch vụ Fiber có số lần báo hỏng cao nhất, tiếp theo là MyTV và Wifi Mesh. Điều này đúng cho cả hai trạng thái rời bỏ và không rời bỏ dịch vụ.
* Biểu đồ đường cho thấy số lần báo hỏng tăng dần qua các năm, đặc biệt từ khoảng năm 2015 trở đi. Đỉnh điểm là vào khoảng năm 2020 và sau đó có xu hướng giảm nhẹ.

Kết luận:

* Mặc dù khách hàng không rời bỏ dịch vụ có số lần báo hỏng cao, nhưng họ vẫn tiếp tục sử dụng dịch vụ. Điều này có thể do sự hài lòng với các khía cạnh khác của dịch vụ hoặc do thiếu lựa chọn thay thế.
* Fiber là loại dịch vụ có số lần báo hỏng cao nhất, cho thấy có thể có vấn đề về chất lượng dịch vụ hoặc khách hàng sử dụng dịch vụ này nhiều hơn. Cần tập trung cải thiện chất lượng dịch vụ Fiber để giảm số lần báo hỏng.
* Số lần báo hỏng tăng dần qua các năm cho thấy có thể có vấn đề về chất lượng dịch vụ hoặc nhu cầu sử dụng dịch vụ tăng lên dẫn đến hệ thống bị quá tải. Cần đầu tư vào nâng cấp và bảo trì hệ thống để giảm thiểu sự cố kỹ thuật.

### **2.4.5 Phân tích khả năng khách hàng rời bỏ theo thời gian sử dụng dịch vụ**



Hình 2.9- Phân tích khả năng khách hàng rời bỏ theo thời gian sử dụng dịch vụ

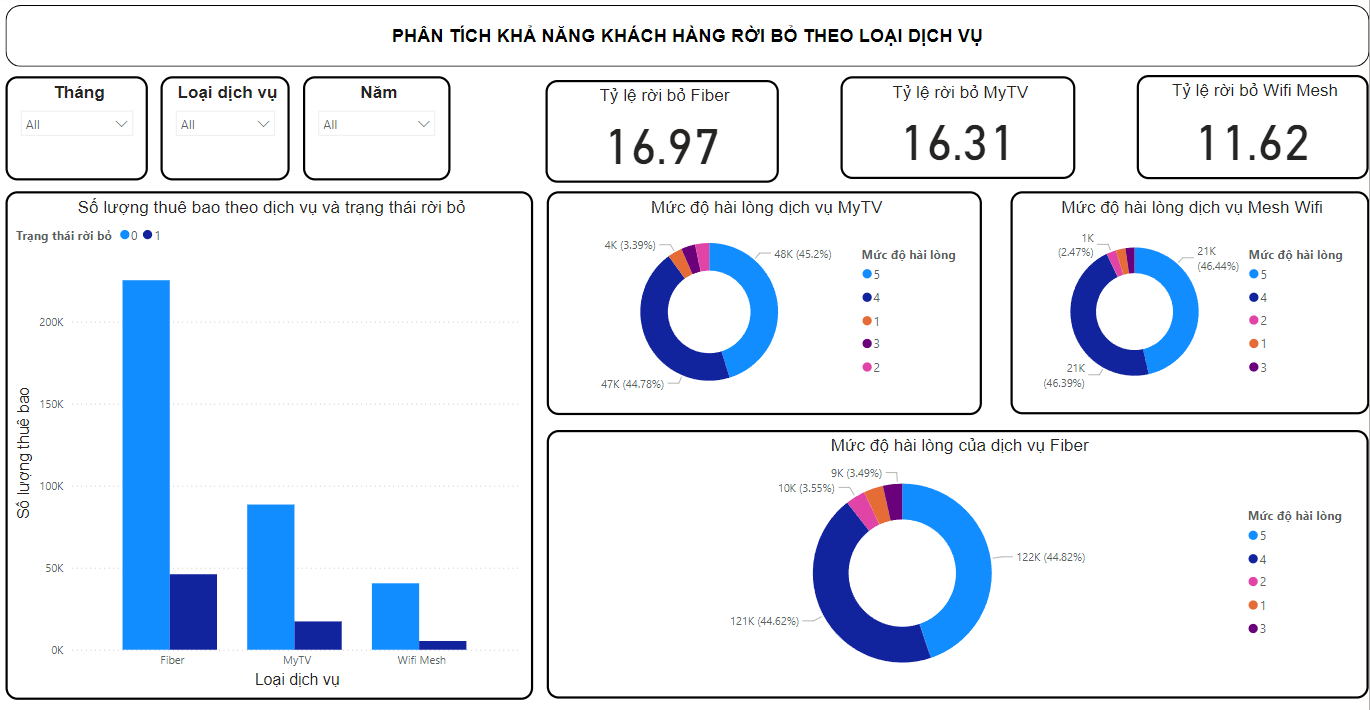
Mô tả:

* Biểu đồ cột bên trái cho thấy số lượng thuê bao rời bỏ dịch vụ tập trung chủ yếu trong năm đầu tiên và năm thứ 2 sử dụng dịch vụ
* Biểu đồ đường cho thấy khách hàng có số tháng sử dụng dịch vụ càng lâu thì càng ít rời bỏ dịch vụ hơn.
* Biểu đồ cột ở dưới cho thấy dịch vụ Fiber có lượng thuê bao lớn nhất trong năm đầu tiên, tiếp theo là MyTV và Wifi Mesh. Các loại dịch vụ này có xu hướng giảm dần sau năm đầu tiên.

Kết luận:

* Phần lớn khách hàng rời bỏ dịch vụ trong năm đầu tiên sử dụng. Điều này có thể chỉ ra rằng khách hàng chưa thấy giá trị hoặc gặp vấn đề về dịch vụ trong giai đoạn đầu sử dụng.Cần có chiến lược chăm sóc khách hàng tốt hơn trong năm đầu tiên để giữ chân họ lâu dài hơn.
* Dịch vụ Fiber có lượng thuê bao lớn nhưng cũng có tỷ lệ rời bỏ cao trong năm đầu tiên. Điều này cho thấy mặc dù khách hàng có nhu cầu cao đối với dịch vụ này, nhưng cũng có thể gặp vấn đề về chất lượng hoặc hỗ trợ kỹ thuật.

**2.4.6 Phân tích khả năng khách hàng rời bỏ theo loại dịch vụ**



Hình 2.10-Phân tích khả năng khách hàng rời bỏ theo loại dịch vụ

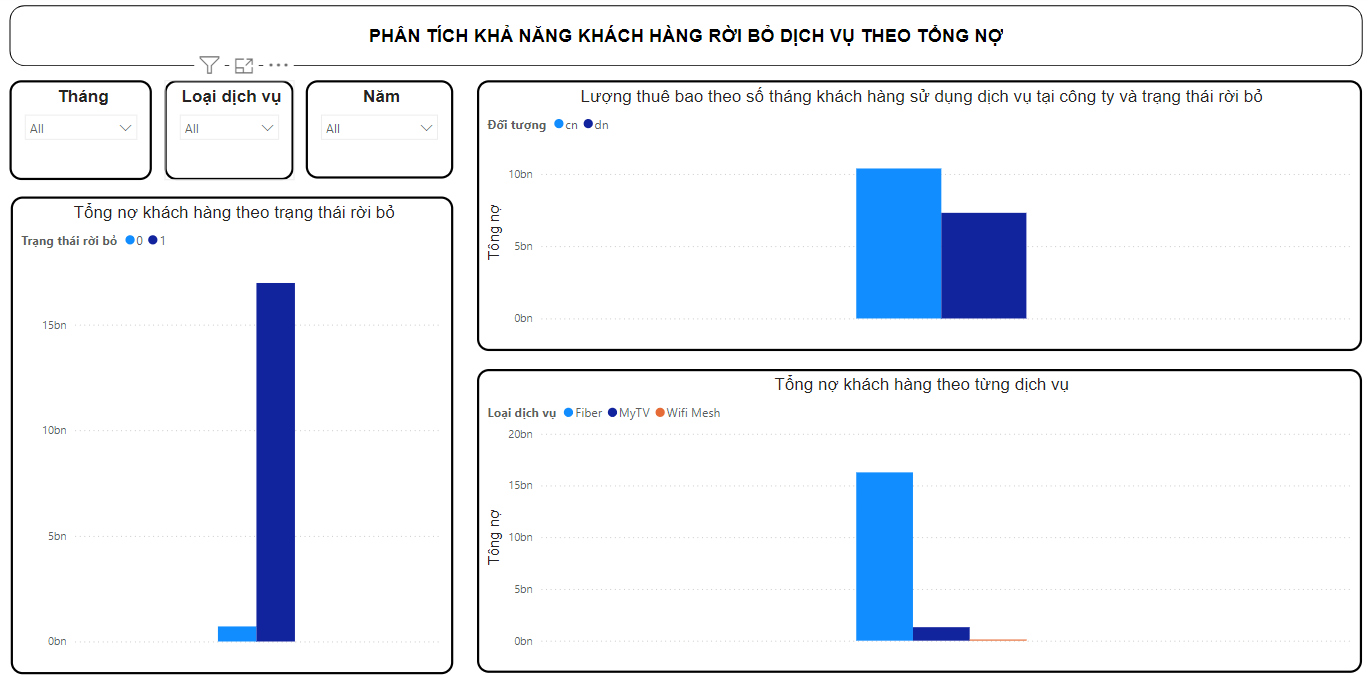
Mô tả:

* Biểu đồ cột cho thấy số lượng thuê bao sử dụng dịch vụ Fiber là cao nhất, tiếp theo là MyTV và Wifi Mesh.
* Số lượng thuê bao rời bỏ (trạng thái 1) của dịch vụ Fiber cũng cao nhất so với các dịch vụ khác.
* Tỷ lệ rời bỏ dịch vụ Fiber là 16.97%, MyTV là 16.31%, và Wifi Mesh là 11.62%.
* Dịch vụ MyTV: Mức độ hài lòng 4 và 5 chiếm tỷ lệ cao nhất, lần lượt là 45.2% và 44.78%.
* Dịch vụ Wifi Mesh: Mức độ hài lòng 4 và 5 chiếm tỷ lệ cao nhất, lần lượt là 46.44% và 46.39%.
* Dịch vụ Fiber: Mức độ hài lòng 4 và 5 chiếm tỷ lệ cao nhất, lần lượt là 44.82% và 44.62%.

Kết luận:

* Dịch vụ Fiber là dịch vụ phổ biến nhất với số lượng thuê bao cao nhất, nhưng cũng có tỷ lệ rời bỏ cao nhất. Điều này có thể chỉ ra rằng mặc dù nhiều khách hàng sử dụng dịch vụ này, nhưng cũng có nhiều khách hàng không hài lòng hoặc gặp vấn đề với dịch vụ.
* Mức độ hài lòng của các dịch vụ đều cao, với phần lớn khách hàng đánh giá mức độ hài lòng ở mức 4 và 5. Tuy nhiên, vẫn có tỷ lệ rời bỏ cao, điều này có thể chỉ ra rằng có những yếu tố khác ngoài mức độ hài lòng ảnh hưởng đến quyết định rời bỏ của khách hàng.
* Wifi Mesh có tỷ lệ rời bỏ thấp nhất (11.62%) so với các dịch vụ khác, cho thấy khách hàng sử dụng dịch vụ này có xu hướng hài lòng và duy trì dịch vụ hơn.

### **2.4.7 Phân tích khả năng khách hàng rời bỏ theo tổng nợ**



Hình 2.11- Phân tích khả năng khách hàng rời bỏ dịch vụ theo tổng nợ

Mô tả:

* Biểu đồ cột bên trái cho thấy tổng nợ của khách hàng rời bỏ dịch vụ (trạng thái 1) rất cao so với tổng nợ của khách hàng không rời bỏ (trạng thái 0).
* Biểu đồ cột bên phải cho thấy số lượng thuê bao của khách hàng không rời bỏ dịch vụ cao hơn so với khách hàng đã rời bỏ dịch vụ. Khách hàng không rời bỏ có tổng nợ cao hơn so với khách hàng rời bỏ.
* Biểu đồ cột dưới cùng cho thấy tổng nợ của khách hàng chủ yếu tập trung ở dịch vụ Fiber, tiếp theo là MyTV và Wifi Mesh.

Kết luận:

* Khách hàng rời bỏ dịch vụ có tổng nợ cao, điều này có thể là do khách hàng không thể thanh toán nợ đúng hạn hoặc không còn khả năng thanh toán. Việc rời bỏ dịch vụ có thể là một biện pháp để tránh nợ.
* Tổng nợ của khách hàng chủ yếu tập trung ở dịch vụ Fiber, cho thấy rằng dịch vụ này có thể có chi phí cao hoặc khách hàng sử dụng nhiều dịch vụ liên quan đến Fiber.

**2.4.8 Phân tích khả năng khách hàng rời bỏ theo trả trước**



Hình 2.12- Phân tích khả năng khách hàng rời bỏ theo trả trước

Mô tả:

* Biểu đồ tròn cho thấy 40.09% khách hàng có trả trước (tương ứng với 169.53K khách hàng) và 59.91% khách hàng không có trả trước (tương ứng với 253.33K khách hàng).
* Biểu đồ cột bên phải cho thấy số lượng thuê bao không trả trước (trạng thái 0) nhiều hơn đáng kể so với số lượng thuê bao trả trước (trạng thái 1). Tuy nhiên, trong nhóm khách hàng rời bỏ, số lượng thuê bao trả trước tương đối thấp.
* Biểu đồ cột ở dưới cùng cho thấy dịch vụ Fiber có số lượng thuê bao trả trước cao nhất, tiếp theo là MyTV và Wifi Mesh.

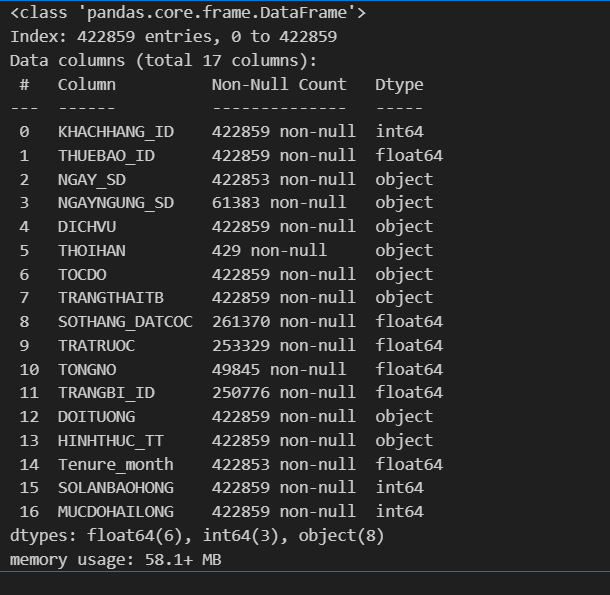
Kết luận:

* Phần lớn khách hàng không có trả trước có xu hướng rời bỏ dịch vụ cao hơn so với khách hàng có trả trước. Điều này có thể chỉ ra rằng chính sách trả trước giúp giảm tỷ lệ rời bỏ dịch vụ.
* Dịch vụ Fiber có số lượng thuê bao trả trước cao nhất, điều này cho thấy khách hàng sử dụng dịch vụ Fiber có xu hướng chọn phương án trả trước nhiều hơn.
* Tỷ lệ khách hàng trả trước chỉ chiếm 40.09%, điều này có thể cho thấy một phần lớn khách hàng không lựa chọn trả trước và có thể gặp khó khăn trong việc thanh toán, dẫn đến rời bỏ dịch vụ.

## 2.5 DỰ ĐOÁN KHÁCH HÀNG RỜI BỎ DỊCH VỤ

### **2.5.1 Xử lý dữ liệu**

Bộ dữ liệu gồm 422859 dòng và 17 cột.



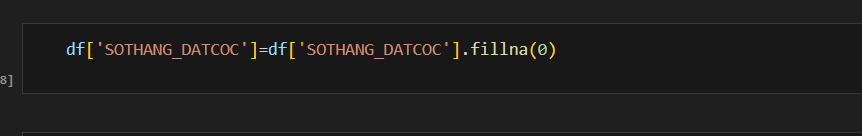
Hình 2.13- Thông tin về bộ dữ liệu

Có những cột không quá cần thiết trong việc dự đoán nên sẽ loại bỏ những cột như: NGAY\_SD,NGAY\_NGUNGSD,KHACHHANG\_ID,THUEBAO\_ID,THOIHAN,TRANGBI\_ID,HINHTHUC\_TT

Một số cột có khá nhiều giá trị NULL bao gồm: SOTHANGDATCOC,TRATRUOC,TONGNO,Tenure\_month ta cần xử lý chúng.

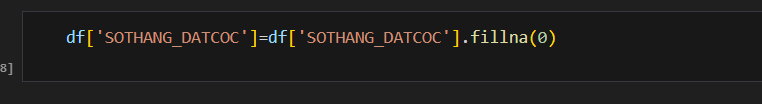
**Với cột SOTHANGDATCOC:**

Điền giá trị là 0 với những dòng bị null.

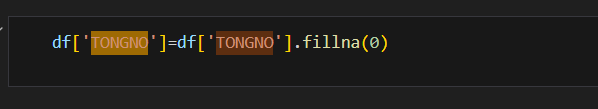


Hình 2.14- Xử lý giá trị null cột SOTHANG\_DATCOC

**Với cột TRATRUOC và TONGNO:**

Điền giá trị 0 với những dòng bị null

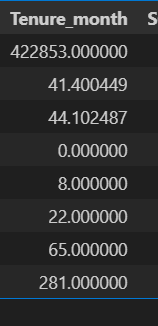
Hình 2.15-Xử lý dữ liệu null cột TRATRUOC



Hình 2.16- Xử lý dữ liệu null với cột TONGNO

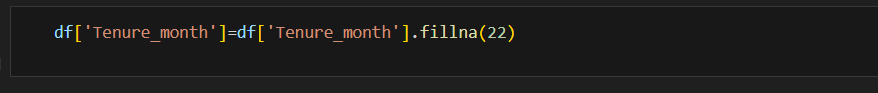
**Với cột tenure\_month:**

Ta lấy giá trị trung bình của tenure\_month điền vào:



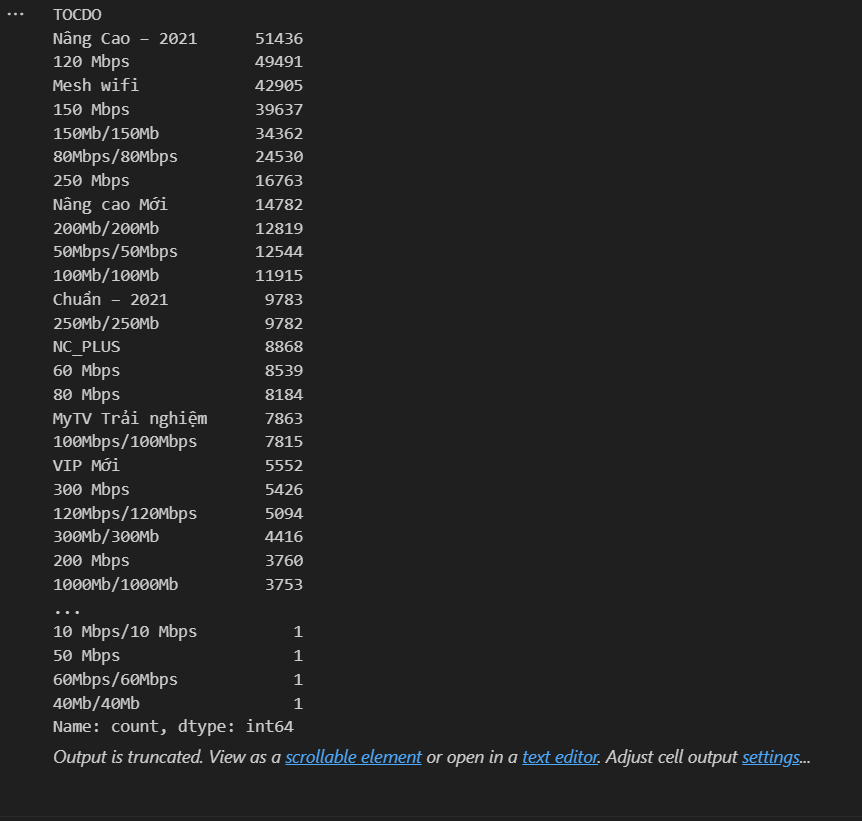
Hình 2.17- Mô tả bộ dữ liệu của Tenure\_month

Ta thấy rằng giá trị trung bình của tenure\_month là 22 nên sẽ điền vào các giá trị bị null.



Hình 2.18- Xử lý giá trị null cột Tenure\_month

**Xử lý cột TOCDO:**



Hình 2.19- Các giá trị của cột TOCDO

Trong cột “TOCDO” có rất nhiều giá trị khác nhau đại diện cho các tốc độ Internet.Các giá trị này có thể được ghi lại dưới nhiều định dạng khác nhau (ví dụ : “100 mbps , 100 mb”,”100 Mpbs/100 Mb”.Bằng cách chuyển đổi tất cả giá trị về dạng chuẩn “X Mpbs”,sẽ giúp cho dữ liệu được nhất quán.

Tiếp đó, có khá nhiều giá trị trong cột không phản ảnh ánh tốc độ của gói cước mà chỉ là gói cước. Do đặc thù của những gói dịch vụ MyTV thường sẽ không có tốc độ trong gói cước. Vì thế với những gói sử dụng dịch vụ MyTV ta quy ước chúng là “0 Mpbs”.

Đối với những gói cước tên là “Mesh Wifi” và “Wifi Mesh 6” do đặc thù của gói cước “Mesh Wifi” thường sẽ có tốc độ là “100 Mbps” còn đối với “Mesh Wifi 6” sẽ có tốc độ từ 300 Mbps đến hơn 800 Mbps. Vì vậy, nên lựa chọn khoảng “500 Mpbs” để đặt giá trị với những dòng là Mesh Wifi 6.

Ngoài ra còn những giá trị ghi là “nan” là những dòng không có thông tin. Tuy nhiên, những dòng đó đều sử dụng dịch vụ Fiber nên ta điền giá trị 100 Mpbs hoặc 150 Mpbs do đặc thù khách hàng sử dụng 2 loại gói cước này khá nhiều.

Sau khi làm cho dữ liệu được nhất quán bộ dữ liệu, để có thể tiến hành phân chia tập huấn luyện và tập kiểm tra ta phải loại bỏ các kí tự chữ cái.Vì thế loại bỏ đuôi “Mpbs” để tiến hành bước phân chia tập huấn luyện và tập kiểm tra.

Đoạn code xử lý dữ liệu:

# Ensure that all TOCDO values are treated as strings

df['TOCDO'] = df['TOCDO'].astype(str).str.strip()

# Replace specific values with '0 Mbps' and others with '500 Mbps' or '150 Mbps'

replacement\_dict = {

    'Basic – 2021': '0 Mbps',

    'Chuẩn': '0 Mbps',

    'Chuẩn Mới': '0 Mbps',

    'Chuẩn – 2021': '0 Mbps',

    'LiveTv HD': '0 Mbps',

    'MYTV Gold HD': '0 Mbps',

    'MyTV Basic': '0 Mbps',

    'MyTV Gold': '0 Mbps',

    'MyTV Silver': '0 Mbps',

    'MyTV Silver HD': '0 Mbps',

    'MyTV Trải nghiệm': '0 Mbps',

    'NC\_PLUS': '0 Mbps',

    'Nâng Cao – 2021': '0 Mbps',

    'Nâng cao': '0 Mbps',

    'Nâng cao Mới': '0 Mbps',

    'VIP': '0 Mbps',

    'VIP Mới': '0 Mbps',

    'Wifi Mesh 6': '500 Mbps',

    'Mesh wifi': '150 Mbps'

}

# Clean up the strings in the TOCDO column before replacing

df['TOCDO'] = df['TOCDO'].replace(replacement\_dict)

# Define the cleaning function

def clean\_tocdo(value):

    if re.search(r'\d+\s\*[Mm][Bb][Pp]?[Ss]?', value):

        match = re.search(r'(\d+)\s\*[Mm][Bb][Pp]?[Ss]?', value)

        if match:

            return match.group(1) + ' Mbps'

    return value.strip()

# Apply the cleaning function to the TOCDO column

df['TOCDO'] = df['TOCDO'].apply(clean\_tocdo)

# Further clean the TOCDO values

def further\_clean\_tocdo(value):

    if re.search(r'\d+\s\*[Mm][Bb][Pp]?[Ss]?', value):

        match = re.search(r'(\d+)', value)

        if match:

            return match.group(1) + ' Mbps'

    return value.strip()

# Apply the further cleaning function to the TOCDO column

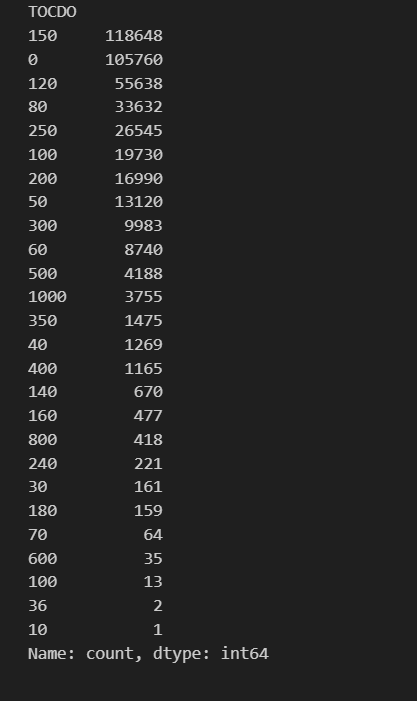
df['TOCDO'] = df['TOCDO'].apply(further\_clean\_tocdo)

# Group by the final cleaned TOCDO values

tocdo\_final\_grouped = df.groupby('TOCDO').size().reset\_index(name='Count')

# Display the result

print(tocdo\_final\_grouped)



Hình 2.20- Kết quả sau khi xử lý cột TOCDO

**Tạo thêm cột TRANGTHAI\_ROIBO làm cột kết quả cho mô hình:**

Ở cột TRANGTHAITHUEBAO nếu các dòng có giá trị “Hoạt động bình thường”,”Tạm dừng”,”Đổi số /Acc” ta đặt biến là 0 (là những khách hàng chưa rời bỏ).Các giá trị còn lại là 1(là những khách hàng đã ngưng sử dụng dịch vụ).

Đoạn code tạo cột TRANGTHAI\_ROIBO:

 # Create the new column TRANGTHAI\_ROIBO

df['TRANGTHAI\_ROIBO'] = df['TRANGTHAITB'].apply(lambda x: 0 if x in ['Hoạt động bình thường', 'Tạm dừng', 'Đổi số/Acc'] else 1)

Sử dụng label\_encoder xử lý cột DOITUONG

Đoạn code:

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler, LabelEncoder,OneHotEncoder

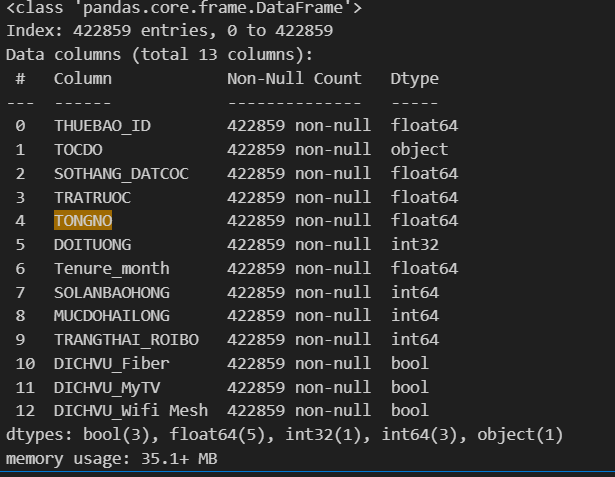
le=LabelEncoder()

df['DOITUONG']=le.fit\_transform(df['DOITUONG'])

Sử dụng onehot-encode cho cột DICHVU

Đoạn code:

df = pd.get\_dummies(df, columns=['DICHVU'])



Hình 2.21 – Kết quả sau khi làm xong các bước xử lý dữ liệu

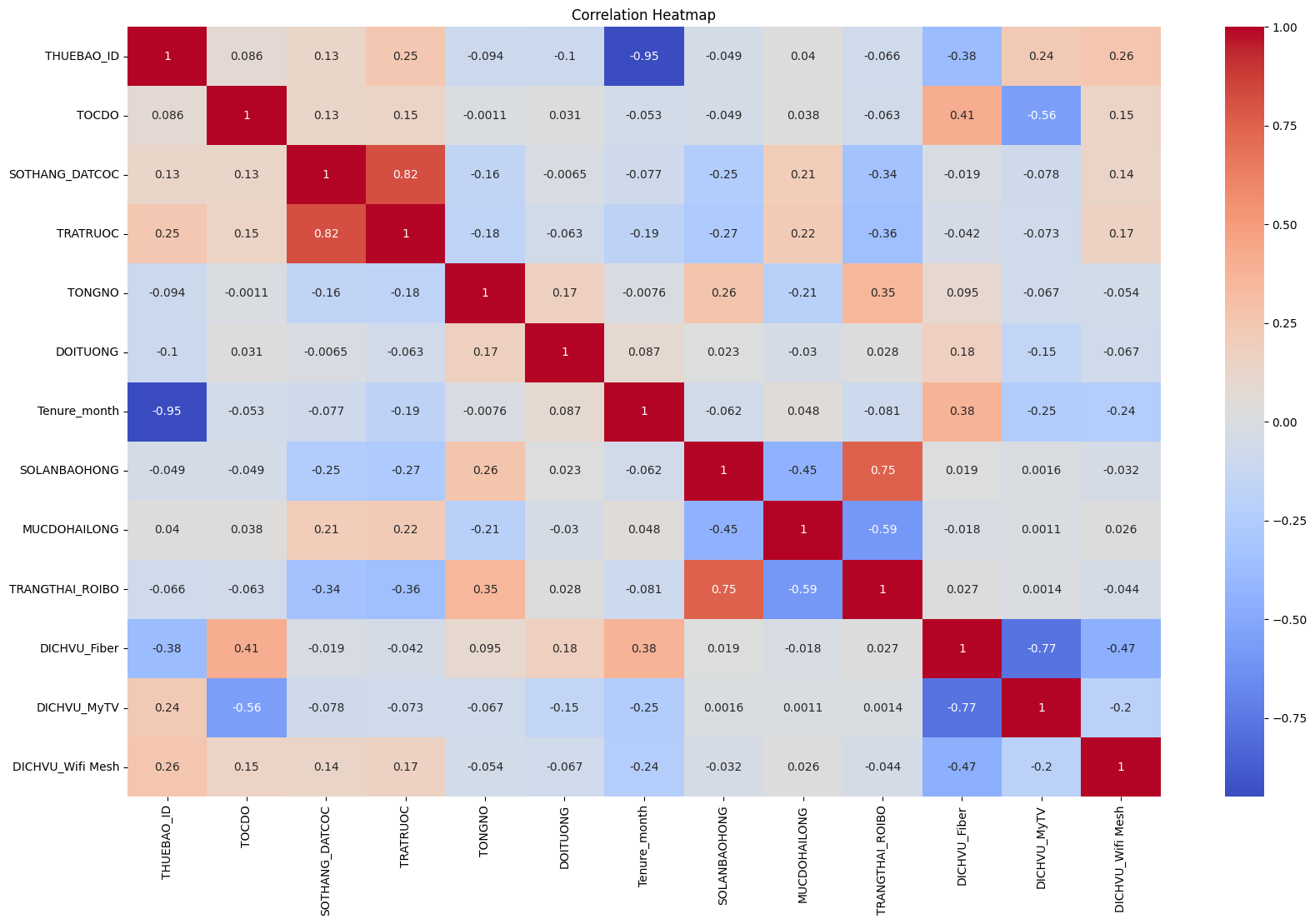
### **2.5.2 Chuẩn hoá bộ dữ liệu**

Loại bỏ các cột không cần thiết cho mô hình:

Đoạn code:

df=df.drop(columns=['HINHTHUC\_TT','NGAY\_SD','THOIHAN','TRANGBI\_ID','KHACHHANG\_ID','TRANGTHAITB'])

Lựa chọn các biến đầu vào (Feature engineer):



Hình 2.22-Biểu đồ mối quan hệ tương quan giữa các biến

Để xây dựng một mô hình đa biến chính xác và đáng tin cậy, cần phải đánh giá mối tương quan giữa các biến giải thích trong mô hình. Nếu một nhóm các biến giải thích có tương quan với nhau, chỉ nên chọn một biến có khả năng phân biệt tốt nhất để đưa vào mô hình cuối cùng khi tiến hành phân tích đa biến:

**+THUEBAO\_ID** và **Tenure\_month** có mối tương quan âm mạnh (-0.95), điều này có nghĩa là khi thời gian sử dụng tăng thì ID thuê bao giảm. Do đó, có thể không cần thiết giữ cả hai biến này trong mô hình.

+**TRATRUOC** và **SOTHANG\_DATCOC** có mối tương quan rất cao (0.82), chỉ nên giữ một trong hai để tránh đa cộng tuyến.

+**SOLANBAOHONG** và **TRANGTHAI\_ROIBO** có mối tương quan dương mạnh (0.75), nghĩa là khi số lần báo hỏng tăng thì trạng thái rời bỏ tăng. Cân nhắc giữ lại một biến này nếu chúng đều có ý nghĩa tương đương.

Dựa trên phân tích trên và theo ngưỡng tương quan 0.7, ta có thể đề xuất các biến cần giữ lại như sau:

**Loại bỏ biến có tương quan cao**:

* Loại bỏ SOTHANG\_DATCOC hoặc TRATRUOC (giữ lại TRATRUOC).

**Giữ lại các biến có ý nghĩa**:

* TOCDO
* TRATRUOC
* TONGNO
* DOITUONG
* Tenure\_month
* SOLANBAOHONG
* MUCDOHAILONG
* TRANGTHAI\_ROIBO ( là biến mục tiêu, không bao gồm trong tập huấn luyện)
* DICHVU\_Fiber
* DICHVU\_Wifi Mesh
* DICHVU\_MyTV

**Phân chia tập huấn luyện:**

-Xác định biến mục tiêu(target) và các biến đầu vào(features):

Đặt biến mục tiêu(Y) là cột TRANGTHAI\_ROIBO

Đặt biến đầu vào (X) là các cột còn lại.

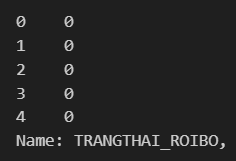
Đoạn code:

X=df.drop(columns=['TRANGTHAI\_ROIBO','THUEBAO\_ID','SOTHANG\_DATCOC'])

Y= df['TRANGTHAI\_ROIBO']

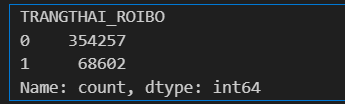


***Hình 2.23- Biến đầu vào(X***)



Hình 2.24- Biến mục tiêu (Y)

**Kiểm tra biến mục tiêu đã cân bằng hay không**



Hình 2.25- Giá trị biến mục tiêu

Do giá trị 0 và giá trị 1 chênh lệch quá nhiều nên cần phải cân bằng lại bộ dữ liệu để cho mô hình được tối ưu và học được tốt.

Vì thế, ta sử dụng over\_sampling để cân bằng lại bộ dữ liệu:

Đoạn code:

from imblearn.over\_sampling import SMOTE

smote = SMOTE()

X, Y = smote.fit\_resample(X, Y)

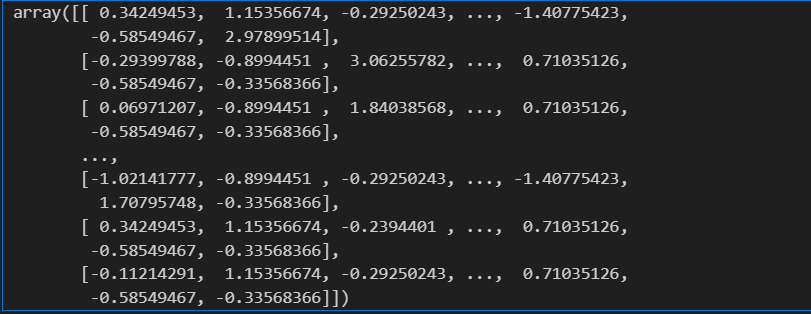
**Chuẩn hoá dữ liệu:**

Standard Scaler chuẩn hóa dữ liệu sao cho mỗi đặc trưng có giá trị trung bình là 0 và độ lệch chuẩn là 1. Công thức chuẩn hóa từng giá trị x như sau:



Trong đó:

* z: là giá trị đã được chuẩn hóa.
* x: là giá trị gốc của đặc trưng.
* μ:là giá trị trung bình của đặc trưng đó trong dữ liệu huấn luyện.
* σ: là độ lệch chuẩn của đặc trưng đó trong dữ liệu huấn luyện.



Hình 2.26-Sử dụng Standard Scaler chuẩn hoá bộ dữ liệu

### **2.5.3 Xây dựng mô hình Machine Learning**

Các mô hình được sử dụng trong bộ dữ liệu như sau:

* Logicstic Regression Classifier
* Nearest Neighbor
* Gradient Boosting Classifier
* Decision Tree Classifier
* Random Forest Classifier
* Neural Network
* Naïve Bayes

Đoạn code:

# Time for Classification Models

import time

from sklearn.decomposition import PCA

from sklearn.linear\_model import LogisticRegression

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn import tree

from sklearn.neural\_network import MLPClassifier

from sklearn.ensemble import GradientBoostingClassifier

from sklearn.gaussian\_process.kernels import RBF

from sklearn.ensemble import RandomForestClassifier

from sklearn.naive\_bayes import GaussianNB

#Khai báo các thuật toán phân lớp sử dụng

dict\_classifiers = {

    "Logistic Regression": LogisticRegression(solver='lbfgs', max\_iter=5000),

    "Nearest Neighbors": KNeighborsClassifier(),

    "Gradient Boosting Classifier": GradientBoostingClassifier(),

    "Decision Tree": tree.DecisionTreeClassifier(),

    "Random Forest": RandomForestClassifier(n\_estimators=18),

    "Neural Net": MLPClassifier(alpha=1),

    "Naive Bayes": GaussianNB(),

}

Sau đó ta xây dựng hàm huấn luyện theo các thuật toán ,xác định thời gian chạy khi huấn luyện các thuật toán.

Đoạn code:

import numpy as np

no\_classifiers = len(dict\_classifiers.keys())

#Xây dựng hàm huấn luyện theo các thuật toán

#Xác định thời gian chạy khi huấn luyện của mỗi thuật toán

#Cho biết Độ chính xác của mô hình khi chạy huấn luyện

#Thời gian chạy huấn luyện của mô hình

def batch\_classify(X\_train\_sc, Y\_train, verbose = True):

    df\_results = pd.DataFrame(data=np.zeros(shape=(no\_classifiers,3)), columns = ['classifier', 'train\_score', 'training\_time'])

    count = 0

    for key, classifier in dict\_classifiers.items():

        t\_start = time.process\_time()

        classifier.fit(X\_train\_sc, Y\_train)

        t\_end = time.process\_time()

        t\_diff = t\_end - t\_start

        train\_score = classifier.score(X\_train\_sc, Y\_train)

        df\_results.loc[count,'classifier'] = key

        df\_results.loc[count,'train\_score'] = train\_score

        df\_results.loc[count,'training\_time'] = t\_diff

        if verbose:

            print("trained {c} in {f:.2f} s".format(c=key, f=t\_diff))

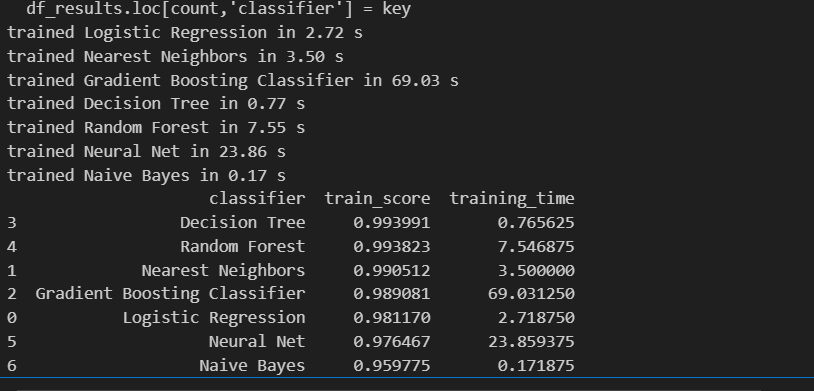
        count+=1

    return df\_results

#Thực hiện việc huấn luyện các mô hình với dữ liệu huấn luyện

df\_results = batch\_classify(X\_train\_sc, y\_train,verbose=True)

print(df\_results.sort\_values(by='train\_score', ascending=False))



Hình 2.27- Kết quả thực hiện mô hình huấn luyện

-Mô hình **Naive Bayes** có thời gian huấn luyện ngắn nhất (0.171875 giây).

-Mô hình **Gradient Boosting Classifier** có thời gian huấn luyện dài nhất (69.031250 giây), điều này phổ biến vì đây là một mô hình phức tạp và mạnh mẽ.

-Các mô hình như **Decision Tree** và **Random Forest** cũng có thời gian huấn luyện tương đối ngắn, cho thấy chúng có thể được huấn luyện nhanh chóng trên tập dữ liệu này.

-Một số mô hình như **Decision Tree**, **Random Forest**, và **Nearest Neighbors** có thể đang bị overfitting với dữ liệu huấn luyện vì độ chính xác quá cao. Cần kiểm tra hiệu suất trên tập dữ liệu kiểm tra (test) để xác nhận điều này.

**-**Các mô hình khác nhau có thời gian huấn luyện khác nhau, và điều này cần được cân nhắc khi lựa chọn mô hình cho các ứng dụng thực tế.

**Áp dụng Cross-Validation Score để giảm tình trạng overfitting:**

Đoạn code:

# Sử dụng  Cross-validation để giải quyết vấn đề overfitting.

#Mỗi thuật toán thực hiện 5 lần huấn luyện trên tập TRAIN với kỹ thuật Cross Validation,

#

#---------------------------------------------------------------

from sklearn.model\_selection import cross\_val\_score

# Logistic Regression

log\_reg = LogisticRegression(solver='lbfgs', max\_iter=5000)

log\_scores = cross\_val\_score(log\_reg, X\_train\_sc, y\_train, cv=5)

log\_reg\_mean = log\_scores.mean()

# KNearestNeighbors

knn\_clf = KNeighborsClassifier()

knn\_scores = cross\_val\_score(knn\_clf, X\_train\_sc, y\_train, cv=5)

knn\_mean = knn\_scores.mean()

# Decision Tree

tree\_clf = tree.DecisionTreeClassifier()

tree\_scores = cross\_val\_score(tree\_clf, X\_train\_sc, y\_train, cv=5)

tree\_mean = tree\_scores.mean()

# Gradient Boosting Classifier

grad\_clf = GradientBoostingClassifier()

grad\_scores = cross\_val\_score(grad\_clf, X\_train\_sc, y\_train, cv=5)

grad\_mean = grad\_scores.mean()

# Random Forest Classifier

rand\_clf = RandomForestClassifier(n\_estimators=18)

rand\_scores = cross\_val\_score(rand\_clf, X\_train\_sc, y\_train, cv=5)

rand\_mean = rand\_scores.mean()

# NeuralNet Classifier

neural\_clf = MLPClassifier(alpha=1)

neural\_scores = cross\_val\_score(neural\_clf, X\_train\_sc, y\_train, cv=5)

neural\_mean = neural\_scores.mean()

# Naives Bayes

nav\_clf = GaussianNB()

nav\_scores = cross\_val\_score(nav\_clf, X\_train\_sc, y\_train, cv=5)

nav\_mean = neural\_scores.mean()

# Create a Dataframe with the results.

d = {'Classifiers': ['Logistic Reg.', 'KNN', 'Dec Tree', 'Grad B CLF', 'Rand FC', 'Neural Classifier', 'Naives Bayes'],

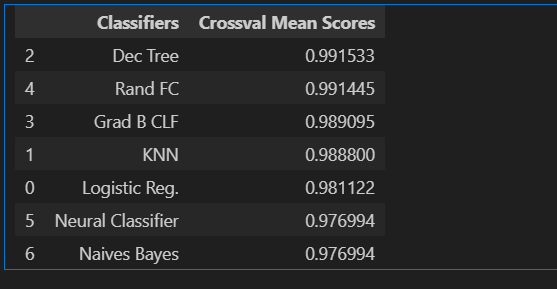
    'Crossval Mean Scores': [log\_reg\_mean, knn\_mean, tree\_mean, grad\_mean, rand\_mean, neural\_mean, nav\_mean]}

result\_df = pd.DataFrame(data=d)

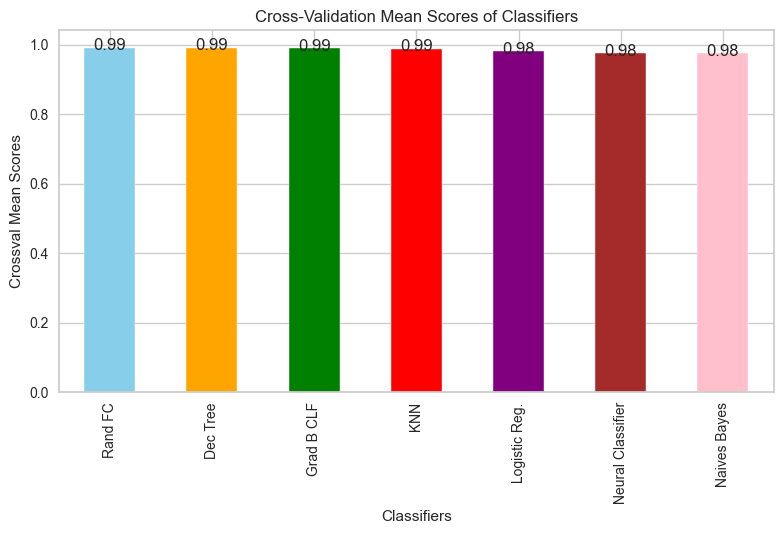
#Hiển thị kết quả của các thuật toán với kỹ thuật Cross Validation

result\_df = result\_df.sort\_values(by=['Crossval Mean Scores'], ascending=False)

result\_df



Hình 2.28 – Kết quả sau khi dùng cross-validation score



Hình 2.29- So sánh độ chính xác của các mô hình

Ta thấy rằng mô hình Decision Tree có điểm số cao nhất, sau đó là mô hình Random Forest và các mô hình còn cũng có điểm số khá cao trên 97% cross-validation score.

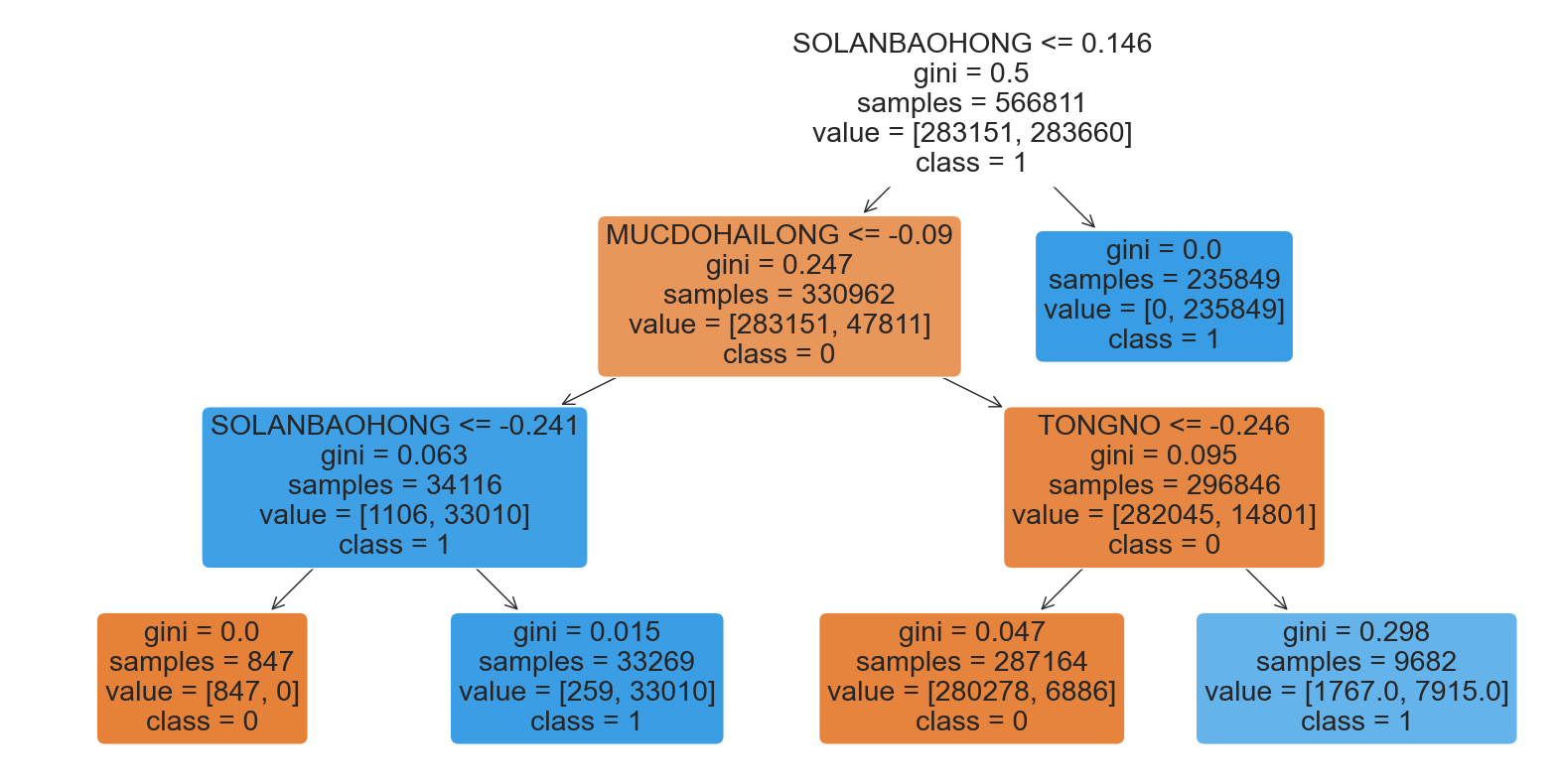
Hình 2.30- So sánh thời gian huấn luyện của các mô hình

Hình 2.30 cho ta thấy thời gian huấn luyện của mô hình. Ta thấy rằng thời gian huấn luyện của mô hình Naïve Bayes và mô hinhg Decision Tree có thời gian huấn luyện thấp nhất so với các mô hình còn lại. Mô hình Gradient Boosting và Neural Network có thời gian huấn luyện cao nhất gấp 20 lần thời gian huấn luyện mô hình Logicstic Regression và KNN(K-nearest neighbors). Và hơn 60 lần so với Naïve Bayes và Decision Tree.

Từ những hình trên ta có thể kết luận mô hình Decision cho ra kết quả huấn luyện tốt nhất và có thời gian huấn luyện ngắn hơn so với các mô hình khác.

### **2.5.4 Giới thiệu về thuật toán Decision Tree**

Mô hình *Decision Tree* là một mô hình được sử dụng khá phổ biến và hiệu quả trong cả hai lớp bài toán phân loại và dự báo của học có giám sát. Khác với những thuật toán khác trong học có giám sát, mô hình *cây quyết định* không tồn tại phương trình dự báo. Mọi việc chúng ta cần thực hiện đó là tìm ra một cây quyết định dự báo tốt trên tập huấn luyện và sử dụng cây quyết định này dự báo trên tập kiểm tra.



Hình 2.31-Mô hình Decision Tree

Cây quyết định (decision tree) trong hình trên là một công cụ phân loại dữ liệu rất hữu ích, giúp chúng ta hiểu rõ quá trình ra quyết định dựa trên các biến đầu vào. Mỗi nút trong cây biểu diễn một đặc điểm (biến), quy tắc quyết định hoặc điều kiện. Chẳng hạn, nút gốc có điều kiện "SOLANBAOHONG <= 0.146" và gini index là 0.5, cho thấy mức độ tạp chất cao nhất có thể. Từ nút gốc này, cây phân chia thành các nhánh, mỗi nhánh đại diện cho kết quả của một quy tắc quyết định và dẫn đến một nút hoặc một lá khác. Các lá biểu thị kết quả cuối cùng hoặc lớp, chứa thông tin về số lượng mẫu trong mỗi lớp.

Việc xử lý dữ liệu như này giúp chúng ta đảm bảo tính nhất quán và độ tin cậy của dữ liệu, từ đó hỗ trợ việc phân tích và báo cáo một cách chính xác hơn. Chẳng hạn, trong cây quyết định này, chúng ta có thể dễ dàng xác định các yếu tố quan trọng ảnh hưởng đến kết quả phân loại, chẳng hạn như "SOLANBAOHONG", "MUCDOHAILONG" và "TONGNO". Điều này giúp tối ưu hóa quá trình ra quyết định và cải thiện hiệu quả phân tích dữ liệu. Cây quyết định không chỉ dễ hiểu mà còn có khả năng bắt kịp các mối quan hệ phi tuyến tính giữa các đặc điểm và biến mục tiêu, làm cho nó trở thành một công cụ mạnh mẽ trong lĩnh vực học máy và phân tích dữ liệu.

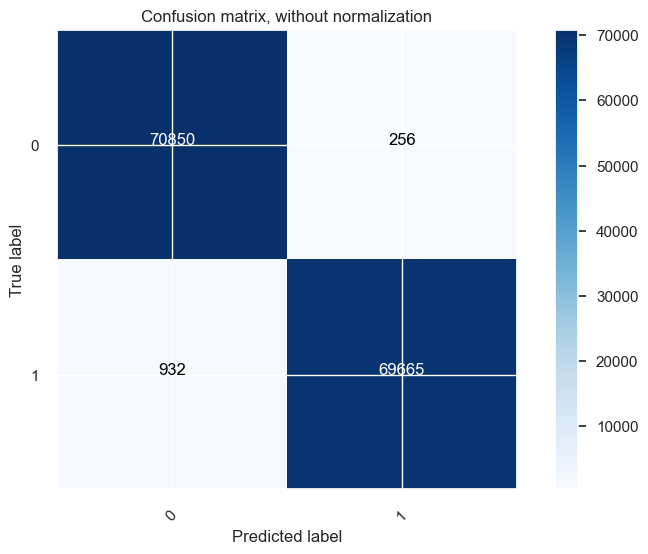
### **2.5.5 Kết quả mô hình Decision Tree**

Ma trận hỗn hợp:

Confusion matrix là một ma trận vuông với kích thước mỗi chiều bằng số lượng lớp dữ liệu. Giá trị tại hàng thứ i, cột thứ j là số lượng điểm lẽ ra thuộc vào class i nhưng lại được dự đoán là thuộc vào class j. Như vậy, nhìn vào confusion matrix (without normalization):

* Hàng (0),cột (0): cho biết số điểm thuộc lớp 0 được phân đúng vào lớp 0 (70850điểm).
* Hàng (0), cột (1): Số điểm thuộc lớp 0 nhưng lại được phân vào lớp 1 (phân lớp sai) 256 điểm.
* Hàng (1), Cột (0): Số điểm thuộc lớp 1 nhưng lại được phân vào lớp 0 (phân lớp sai) 932 điểm
* Hàng (1), cột (1): Số điểm thuộc lớp 1 được phân đúng vào lớp 1 (69665 điểm)

Chúng ta có thể suy ra ngay rằng tổng các phần tử trong toàn ma trận này chính là số điểm trong tập TEST. Các phần tử trên đường chéo của ma trận là số điểm được phân loại đúng của mỗi lớp dữ liệu. Từ đây có thể suy ra accuracy chính bằng tổng các phần tử trên đường chéo chia cho tổng các phần tử của toàn ma trận.



Hình 2.32- Ma trận hỗn hợpChart, pie chart

Description automatically generated

Hình 2.33- Công thức tính điểm Precision và recall trong bài toán phân loại

Với cách định nghĩa một lớp là dương, Độ chính xác được định nghĩa là tỷ lệ giữa số lượng dương thực sự trong số những dương tính được phân loại là dương (TP + FP). Ngoài ra, Precision là số 0 âm nhỏ hơn hoặc bằng một.

A picture containing diagram

Description automatically generated

Khi độ chính xác = 1, tất cả các điểm được tìm thấy đều thực sự dương, nghĩa là không có điểm âm nào được trộn vào kết quả. Tuy nhiên, Precision = 1 không đảm bảo rằng mô hình là tốt, vì câu hỏi đặt ra là liệu mô hình có tìm thấy tất cả các mặt tích cực hay không. Nếu một mô hình chỉ tìm thấy một điểm tích cực mà nó chắc chắn nhất thì nó không thể được gọi là mô hình tốt.

TN: 70850

FN:932

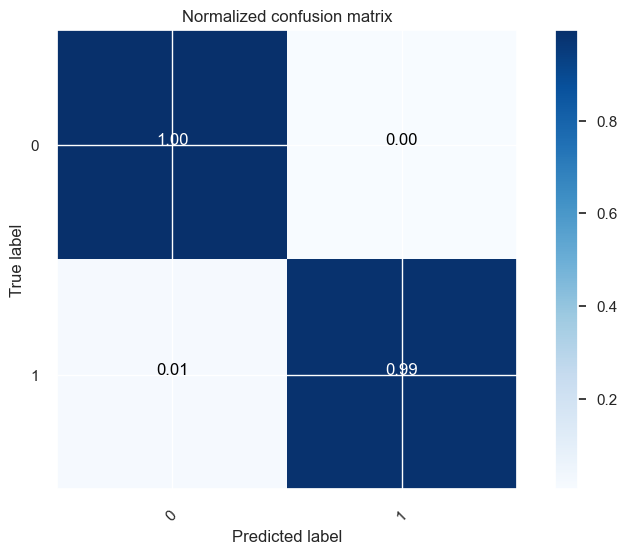
TP:69665

FP:256

Trong báo cáo này, độ chính xác sẽ là:

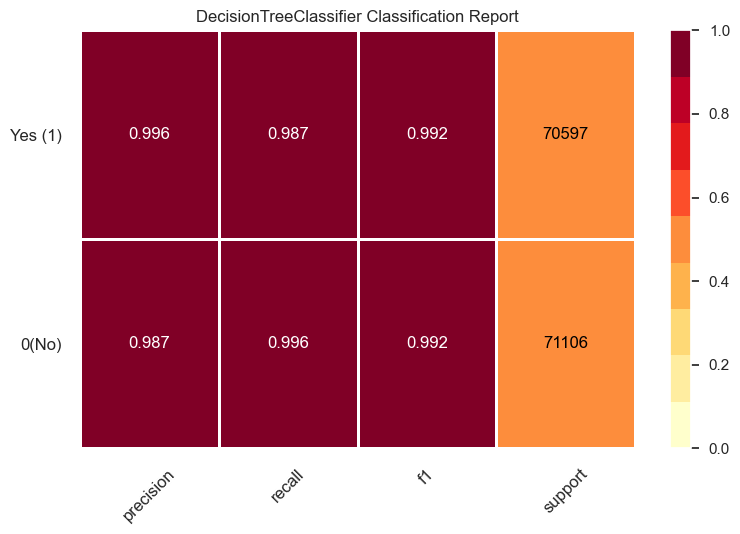
Precision = = 0.9963

Recall = = 0.9867



Hình 2.34- Ma trận nhầm lẫn chuẩn hóa

Để có cái nhìn rõ hơn, ta có thể dùng normalized confusion matrix, tức ma trận hỗn hợp được chuẩn hoá. Để có normalized confusion matrix, ta lấy mỗi hàng của unnormalized confusion matrix sẽ được chia cho tổng các phần tử trên hàng đó. Như vậy, ta có nhận xét rằng tổng các phần tử trên một hàng của normalized confusion matrix luôn bằng 1. Điều này thường không đúng trên mỗi cột. Ma trận hôn hợp thường được minh hoạ bằng màu sắc để có cái nhìn rõ ràng hơn. Với các bài toán với nhiều lớp dữ liệu, cách biểu diễn bằng màu này rất hữu ích. Các ô màu đậm thể hiện các giá trị cao. Một mô hình tốt sẽ cho một confusion matrix có các phần tử trên đường chéo chính có giá trị lớn, các phần tử còn lại có giá trị nhỏ.



Hình 2.35- Bảng báo cáo phân loại

Báo cáo phân loại hiển thị bản trình bày các số liệu phân loại chính trên cơ sở từng lớp. Điều này mang lại trực giác sâu sắc hơn về hành vi của bộ phân loại đối với độ chính xác tổng thể, điều này có thể che giấu những điểm yếu về chức năng trong một lớp của bài toán nhiều lớp. Báo cáo phân loại trực quan được sử dụng để so sánh các mô hình phân loại với các mô hình chọn lọc “đỏ hơn”, ví dụ: có số liệu phân loại mạnh hơn hoặc cân bằng hơn.

Các số liệu được xác định theo kết quả dương tính đúng và sai cũng như âm tính đúng và sai. Tích cực và tiêu cực trong trường hợp này là tên chung cho các lớp của bài toán phân loại nhị phân. Trong ví dụ trên, chúng ta sẽ xem xét có khách hàng rời dịch vụ và không rời dịch vụ. Do đó, giá trị dương thực sự là khi lớp thực tế cũng dương như lớp ước tính. Kết quả dương tính giả là khi lớp thực tế âm nhưng lớp ước tính là dương. Sử dụng thuật ngữ này, các thước đo được định nghĩa như sau:

**Precision:**

Độ chính xác là khả năng của người phân loại không gắn nhãn một trường hợp dương tính mà thực tế là âm tính. Đối với mỗi lớp, nó được định nghĩa là tỷ lệ của kết quả dương tính đúng với tổng số kết quả dương tính đúng và sai. Nói cách khác, “đối với tất cả các trường hợp được phân loại là dương tính, bao nhiêu phần trăm là đúng?”

**Recall:**

Nhớ lại là khả năng của một bộ phân loại để tìm thấy tất cả các trường hợp tích cực. Đối với mỗi loại, nó được định nghĩa là tỷ lệ của kết quả dương tính thực với tổng số kết quả dương tính thực và âm tính giả. Nói cách khác, “đối với tất cả các trường hợp thực sự dương tính, bao nhiêu phần trăm được phân loại chính xác?”

**F1-Score:**

Điểm F1 là giá trị trung bình hài hòa có trọng số của độ chính xác và khả năng thu hồi sao cho điểm tốt nhất là 1,0 và điểm tệ nhất là 0,0. Nói chung, điểm F1 thấp hơn các thước đo độ chính xác vì chúng đưa độ chính xác và khả năng thu hồi vào tính toán của chúng. Theo nguyên tắc chung, nên sử dụng giá trị trung bình có trọng số của F1 để so sánh các mô hình phân loại, chứ không phải độ chính xác toàn cầu.

**Support**

Hỗ trợ là số lần xuất hiện thực tế của lớp trong tập dữ liệu đã chỉ định. Hỗ trợ không cân bằng trong dữ liệu huấn luyện có thể chỉ ra điểm yếu về cấu trúc trong điểm số được báo cáo của bộ phân loại và có thể cho thấy sự cần thiết phải lấy mẫu phân tầng hoặc tái cân bằng. Hỗ trợ không thay đổi giữa các mô hình mà thay vào đó chẩn đoán quá trình đánh giá.

**Confusion Matrix:**

Mục đích chính của Confusion Matrix là để xem mô hình của ta hoạt động như thế nào khi phân loại khách hàng có khả năng rời dịch vụ viễn thông. Chúng ta sẽ thấy trong ma trận nhầm lẫn bốn thuật ngữ là Dương tính thật, Dương tính giả, Âm tính thật và Âm tính giả.

Positive/Negative: Loại Lớp (nhãn) ["No", "Yes"]

True/False: Phân loại đúng hoặc sai theo mô hình.

**True Negatives (TN)**: Phân loại chính xác loại "Không" hoặc khách hàng không rời dịch vụ viễn thông.

**False Positives (FP)**: Phân loại không chính xác loại "Có" trong khi thực tế là "Không" (khách hàng không rời dịch vụ nhưng bị dự đoán là rời).

**False Negatives (FN)**: Phân loại không chính xác loại "Không" trong khi thực tế là "Có" (khách hàng rời dịch vụ nhưng bị dự đoán là không rời).

**True Positives (TP)**: Phân loại chính xác loại "Có" hoặc khách hàng rời dịch vụ viễn thông.

**Phân Tích Báo Cáo Phân Loại:**

**Lớp Yes (1):**

- **Precision**: 0.996 (99.6%)

  Trong số các mẫu được dự đoán là "Yes", 99.6% là chính xác.

- **Recall**: 0.987 (98.7%)

  Trong số các mẫu thực sự là "Yes", mô hình dự đoán đúng 98.7%.

- **F1 Score**: 0.992 (99.2%)

Sự cân bằng giữa precision và recall là 99.2%.

- **Support**: 70597

 Có 70597 mẫu thực sự thuộc lớp "Yes" trong tập dữ liệu kiểm tra.

**Lớp No (0):**

- **Precision**: 0.987 (98.7%)

  Trong số các mẫu được dự đoán là "No", 98.7% là chính xác.

- **Recall**: 0.996 (99.6%)

  Trong số các mẫu thực sự là "No", mô hình dự đoán đúng 99.6%

- **F1 Score**: 0.992 (99.2%)

  Sự cân bằng giữa precision và recall là 81.7%.

- **Support**: 71106

  - Có 999 mẫu thực sự thuộc lớp "No" trong tập dữ liệu kiểm tra.

**Kết Luận**:

**Precision** cao hơn cho lớp "No" chỉ ra rằng mô hình ít khi dự đoán nhầm lớp "No" hơn.

**Recall** cao hơn cho lớp "Yes" chỉ ra rằng mô hình có khả năng tìm thấy hầu hết các trường hợp thực sự là "Yes".

**F1 Score** cho cả hai lớp đều cao, cho thấy mô hình có sự cân bằng tốt giữa precision và recall.

**Support** cho thấy số lượng mẫu thực tế của mỗi lớp, giúp hiểu rõ hơn về dữ liệu huấn luyện.

Tóm lại, báo cáo phân loại và ma trận nhầm lẫn cung cấp cái nhìn sâu sắc về hiệu suất của mô hình trên từng lớp, giúp xác định các điểm mạnh và yếu của mô hình, từ đó đưa ra các điều chỉnh cần thiết để cải thiện hiệu suất phân loại.

### **2.5.6 Kết luận**

Mô hình Decision Tree cho thấy tính nhất quán và độ tin cậy cao trong việc phân loại khách hàng rời bỏ và không rời bỏ dịch vụ. Các chỉ số precision, recall và F1 score cao đồng nghĩa với việc mô hình ít mắc lỗi và có khả năng phát hiện chính xác các trường hợp khách hàng rời bỏ dịch vụ.

Với precision và recall cao, mô hình có khả năng phát hiện chính xác các trường hợp khách hàng thực sự rời bỏ dịch vụ mà không gán nhãn nhầm cho những khách hàng không rời bỏ. Điều này quan trọng trong việc giữ chân khách hàng và đưa ra các biện pháp cải thiện dịch vụ.

F1 score cao cho cả hai lớp cho thấy mô hình có sự cân bằng tốt giữa precision và recall, nghĩa là mô hình không chỉ chính xác trong việc phân loại mà còn có khả năng phát hiện đầy đủ các trường hợp thực sự dương tính và âm tính.

Số lượng mẫu thực tế của mỗi lớp (support) cho thấy mô hình được huấn luyện trên dữ liệu khá cân bằng, giúp cải thiện hiệu suất của mô hình và tránh việc mô hình bị thiên lệch về một lớp nào đó.

### **2.5.7 Đề xuất cải tiến**

Nếu có thêm dữ liệu, đặc biệt là dữ liệu liên quan đến hành vi khách hàng và các yếu tố khác, mô hình có thể được cải thiện hơn nữa.

Liên tục kiểm tra và điều chỉnh mô hình dựa trên dữ liệu mới để đảm bảo mô hình luôn phản ánh chính xác hành vi của khách hàng và các yếu tố ảnh hưởng đến quyết định rời bỏ dịch vụ.

# CHƯƠNG 3. ĐÁNH GIÁ KẾT QUẢ VÀ HƯỚNG PHÁT TRIỂN

## 3.1. ĐÁNH GIÁ KẾT QUẢ

### **3.1.1 Kết quả đã làm được**

Trong quá trình thực tập ở đơn vị em đã làm được những công việc như sau:

* Tìm hiểu về bộ dữ liệu, các trường thông tin dữ liệu,các công cụ trong phân tích dữ liệu.
* Xử lý và làm sạch bộ dữ liệu.
* Khai phá bộ dữ liệu.
* Phân tích các trường thông tin và trực quan hoá dữ liệu.
* Tạo bảng báo cáo (dashboard).
* Xây dựng mô hình dự đoán khách hàng rời bỏ dịch vụ.
* Đánh giá thuật toán bằng các chỉ số như độ chính xác, độ phủ, độ đặc hiệu ,nhằm chọn ra mô hình tốt nhất.

Trong quá trình thực tập tại đơn vị, các công việc mà em đã thực hiện mang lại nhiều kết quả đáng kể. Đầu tiên, qua việc tìm hiểu bộ dữ liệu và các trường thông tin, đơn vị có thể xác định rõ cấu trúc và nội dung của dữ liệu, giúp dễ dàng trong việc quản lý và sử dụng thông tin. Việc xử lý và làm sạch dữ liệu đảm bảo loại bỏ những giá trị sai lệch, thiếu hoặc không hợp lý, từ đó nâng cao độ tin cậy và tính chính xác của dữ liệu.

Khai phá dữ liệu đã giúp phát hiện ra những mẫu hình và xu hướng tiềm ẩn, cung cấp những thông tin giá trị về hành vi và đặc điểm của khách hàng.

Phân tích các trường thông tin và trực quan hóa dữ liệu thông qua biểu đồ và đồ thị giúp đơn vị có cái nhìn trực quan, dễ hiểu về các yếu tố quan trọng ảnh hưởng đến quyết định rời bỏ dịch vụ của khách hàng. Việc tạo bảng báo cáo (dashboard) giúp đơn vị dễ dàng theo dõi và giám sát các chỉ số quan trọng theo thời gian thực, hỗ trợ quá trình ra quyết định chiến lược một cách nhanh chóng và hiệu quả.

Xây dựng mô hình dự đoán khách hàng rời bỏ dịch vụ, cùng với việc đánh giá các thuật toán bằng các chỉ số như độ chính xác, độ phủ, độ đặc hiệu, giúp đơn vị chọn ra mô hình tốt nhất, từ đó dự đoán chính xác hơn về nguy cơ mất khách hàng.

Tổng hợp lại, những công việc này đã cung cấp cho đơn vị một cái nhìn toàn diện về tình trạng khách hàng, giúp nhận diện kịp thời các dấu hiệu rời bỏ và đưa ra các chiến lược giữ chân khách hàng hiệu quả, đồng thời tối ưu hóa các hoạt động kinh doanh và nâng cao chất lượng dịch vụ.

### **3.1.2 Hạn chế**

Do một số nguyên nhân khách quan và chủ quan cũng như việc hạn chế về mặt thời gian mà đề tài vẫn chưa hoàn thiện. Dữ liệu thức tế có sự chênh lệch khá lớn giữa thuê bao đã rời bỏ và thuê bao chưa rời bỏ cũng như chưa bao quát được hết các trường hợp.Dữ liệu mẫu cần được huấn luyện và mở rộng môi trường áp dụng.

Các trường hợp phân loại sai vẫn còn nhiều dẫn đến việc nhằm mục tiêu khách hàng có nguy cơ cao chưa thật sự chuẩn xác.

Mô hình dự đoán chưa phân tích sâu vào các tham số để phù hợp với mô hình dữ liệu thực tế

## 3.2 HƯỚNG PHÁT TRIỂN

Tập trung nghiên cứu rút trích các đặc trưng thuộc tính phù hợp hơn cho quá trình phân tích, tăng độ chính xác trong việc dự đoán tập khách hàng có nguy cơ cao.

Nghiên cứu các mô hình dự đoán để cải thiện mô hình dự đoán được tốt hơn. Nghiên cứu áp dụng các mô hình phân loại kết hợp để tìm kiếm những mô hình tối ưu phù hợp với dữ liệu thực tế tại đơn vị.

Tiến hành áp dụng tại doanh nghiệp. Cảnh báo sớm các nhóm khách hàng có nguy cơ cao, góp phần hỗ trợ công tác chăm sóc và lôi kéo khách hàng được tiến hành nhanh và hiệu quả hơn. Từ đó, góp phần thúc đẩy hiệu quả kinh doanh tại đơn vị.

## 3.3 KẾT LUẬN

Sau khoảng 2 tháng thực tập tại Công ty Viễn Thông Bình Dương VNPT, em đã có cơ hội học hỏi và trải nghiệm thực tế rất nhiều điều bổ ích. Qua quá trình làm việc tại đây, em đã có thể áp dụng những kiến thức lý thuyết đã học vào thực tiễn, từ đó củng cố và mở rộng hiểu biết của mình về ngành viễn thông.

Trong thời gian thực tập, em đã hoàn thành các công việc quan trọng như xử lý và làm sạch dữ liệu, khai phá dữ liệu, phân tích và trực quan hoá dữ liệu, tạo bảng báo cáo (dashboard), xây dựng và đánh giá mô hình dự đoán khách hàng rời bỏ dịch vụ và được tham gia vào dự án “Phân tích và dự đoán khách rời bỏ dịch vụ trong nghành viễn thông”.

Thông qua quá trình thực tập, em đã hiểu rõ hơn về môi trường làm việc chuyên nghiệp, quy trình làm việc trong ngành viễn thông, và những thách thức cũng như cơ hội trong lĩnh vực này. Em tin rằng những kinh nghiệm và kỹ năng tích lũy được trong thời gian thực tập sẽ là nền tảng vững chắc giúp em tự tin bước vào sự nghiệp sau khi tốt nghiệp.

Em xin chân thành cảm ơn sự hướng dẫn và hỗ trợ tận tình từ các anh chị đồng nghiệp và ban lãnh đạo công ty trong suốt thời gian thực tập. Những kiến thức và kinh nghiệm mà em đã học hỏi được tại Công ty Viễn Thông Bình Dương VNPT sẽ là hành trang quý báu cho em trên con đường phát triển nghề nghiệp trong tương lai.

# TÀI LIỆU THAM KHẢO

Tiếng Việt

|  |  |
| --- | --- |
| [1] | P. H. Hoàng, Machine Learning Cơ Bản, Hà Nội: Nhà xuất bản Khoa học và Kỹ thuật, 2020. |
| [2] | H. V. T. Phong, Kể Chuyện Thông Qua Dữ Liệu - Cole Nussbaumer Knaflic, Hà Nội: Nhà Xuất Bản Thế Giới, 2021. |
| [3] | Alistair Croll,Benjamin Yoskovitz, Phân Tích Dữ Liệu Tinh Gọn, Hồ Chí Minh: Nhà xuất bản Trẻ, 2014. |
| [4] | T. Q. Hải, Thống Kê Ứng Dụng trong Phân Tích Dữ Liệu Kinh Doanh, Hà Nội: Nhà xuất bản Tài chính, 2018. |

Tiếng Anh

|  |  |
| --- | --- |
| [1] | A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, Sebastopol, CA, USA: O'Reilly Media, 2019. |
| [2] | K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, MA, USA: MIT Press, 2012. |
| [3] | Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, New York, NY, USA: Springer, 2009. |